Aalto University
School of Science

Chalmers University of Technology
Department of Mathematical Sciences

Nordic Five Tech MSc Programme in Applied and Engineering Mathematics

Natalia Vesselinova

Large-scale Statistical Inference in Internet of Things Scenarios

Master's Thesis
Espoo, June 11, 2019

Supervisors: Assistant Professor Pauliina Ilmonen, Aalto University
Professor Rebecka Jörnsten, Chalmers University of Technology

Aalto University
School of Science

Chalmers University of Technology
Department of Mathematical Sciences

Nordic Five Tech MSc Programme in Applied and Engineering Mathematics

ABSTRACT OF MASTER'S THESIS

Author: Natalia Vesselinova

Title: Large-scale Statistical Inference in Internet of Things Scenarios

Date: June 11, 2019 **Pages:** vii + 84

Supervisors: Assistant Professor Pauliina Ilmonen
Professor Rebecka Jörnsten

The problem we address is performing statistical inference in Internet of Things scenarios, which are typically composed by a massive number of sensor nodes. We consider a traditional distributed scenario in which sensors make local observations about monitored phenomena and supply a central node (fusion centre) with relevant statistics for making a global decision about the state of the network. The primary contribution of the proposed statistical inference framework is that it is tailored for large-scale multisensor networks by combining non-parametric local detection approach with multiple hypotheses testing procedure that controls error rates. In particular, a local detector is employed by each sensor node. The decision problem at the detector is formulated as a binary hypothesis. The detector employs bootstrapping and non-parametric two-sample Anderson-Darling test to approximate the probability function of and calculate relevant test statistics. The fusion centre employs the False Discovery Rate control procedure for simultaneously evaluating the massive number of sensors test statistics as well as for controlling error rates. For a large number of sensors, demanding local conditions (probability models that resemble each other under the null and alternative hypotheses) and several sensors observing departure from nominal conditions, the power of the proposed statistical framework is large (above 90%). When only few events occur, it might be challenging for the fusion centre to detect them all. To understand and address this problem we analysed the simulation results, which brought insights into the performance of the proposed statistical inference approach as well as potential ways of improving it.

Keywords: statistical inference, non-parametric distributed detection, multiple hypotheses testing, false discovery rate (FDR) control, Internet of Things (IoT), wireless sensor networks

Language: English

Acknowledgements

The research reported in the thesis was conducted during the summer of 2017 at Aalto University under the WiFiUS Project: Secure Inference in the Internet of Things with PI Professor Visa Koivunen, Signal Processing and Acoustics Department, Aalto University. I thank Prof. Koivunen for the opportunity to collaborate with him on the interesting and relevant IoT topic. The computational resources were provided by the Aalto Science-IT project.

I much appreciate the possibility to learn from, the meaningful discussions with and scientific advice of my supervisors Professor Rebecka Jörnsten from the Department of Mathematical Sciences at Chalmers University of Technology and Assistant Professor Pauliina Ilmonen from the Department of Mathematics and Systems Analysis at Aalto University—two scientists, who are impactful lecturers, passionate statisticians and inspirational women with their life accomplishments.

I am grateful to Anu Kuusela, the coordinator of the study programme, for her professional and prompt assistance in solving administrative issues.

This educational journey in mathematics has been a dear aspiration of mine. My heartfelt gratitude to Professor Jarmo Harju from Tampere University of Technology for supporting my application to the mathematical track as well as for his professional support and relevant scientific advice throughout my career. Transparency, respect, fairness, trust and openness were a day-to-day practice at Jarmo's research group. Kiitos paljon Jarmo!

I thank my sister Iva, mother Ivanka, and father Vasil for their endless faith in me and love, one facet of which is their unequivocal support in all my endeavours. I am blessed to be part of the family I am part of including my grandparents, relatives and my predecessors, who have instilled homespun values in us that have been making the life journey beautiful and fascinating, where challenges are seen as teachers and life is looked from its brightest side.

Abbreviations and acronyms

3GPP	3rd Generation Partnership Project
BCI	bootstrap confidence intervals
CDF	cumulative distribution function
CI	confidence interval
EDF	empirical (cumulative) distribution function
FDP	false discovery proportion
FDR	false discovery rate
FWER	family wise error rate
IoT	Internet of Things
MC	Monte Carlo
PCER	per-comparison error rate
PFER	per-family error rate
SNR	signal-to-noise ratio

Mathematical symbols

l	size of observation sample Y
m	total number of hypotheses
m_0	total number of true null hypotheses
m_1	total number of true non-null (alternative) hypotheses
n	size of ambient sample X (under nominal conditions)
p_d	detection probability
μ	mean
σ	standard deviation
σ^2	variance
π_0	proportion of true null hypotheses
π_1	proportion of true non-null (alternative) hypotheses
Φ	cumulative distribution function of the standard normal distribution
q^*	false discovery rate threshold
$\mathcal{N}(0,1)$	standard normal distribution
Q	false discovery proportion
Q_e	false discovery rate
$\mathcal{H}_0$	null hypothesis
$\mathcal{H}_1$	non-null (alternative) hypothesis
F	probability model under nominal conditions
G	probability distribution, which represents conditions under the alternative hypothesis
R	total number of discoveries
S	total number of true discoveries
T	total number of type II errors (missed detections)
U	total number of non-rejected true null hypotheses
V	total number of type I errors (false discoveries)
X	ambient sample obtained under nominal conditions
Y	observation sample

Contents

Chapter 1

Introduction

The Internet of Things (IoT) paradigm is aimed at solving challenges ranging from medical (such as ensuring remote healthcare) and environmental (such as natural disaster prevention) to public (such as indoor safety monitoring). IoT is underlying the "smart world" concept, which encompasses "smart city" (for improved quality of life in urban conglomerations by providing improved public services and environmental conditions for instance), "smart building" (for improved safety and control of heating, lightning and air conditioning to name a few) and "smart agriculture" (for optimised performance and reduced impact on nature) among others. Some practical IoT examples include traffic light control based on current and anticipated number of vehicles on the road to prevent rush hour chaos in large cities [46] or efficient collection of waste in urban and rural areas.

In technical terms IoT refers to smart, autonomous and heterogeneous devices, which can sense physical phenomena, process data and communicate with each other and/or the environment. The IoT has evolved from the Internet and wireless sensor networks and therefore shares important similarities with these two technologies. The most prominent feature of the former is its ubiquitous presence and of the latter: low-cost, low-power, small-size and multifunctional sensor devices [41], [42], which comprise a large, multinodal network through which an area of interest is monitored. Several wireless low power wide area network (WAN) communication technologies have emerged to make possible the practical implementation of the IoT concept. 3GPP, for instance, has developed Narrowband Internet of Things (NB-IoT), LTE machine type communication (LTE-M), and the extended coverage (EC-GSM-IoT) standards. LoRa is a proprietary long-range WAN wireless protocol designed for IoT too.

1.1 Problem statement

We model a general Internet of Things (IoT) scenario by focusing on the underlying sensor network. We make minimum assumptions about the network structure. In particular, the distribution of the sensor nodes—the two extremes being a regular grid of sensors and a randomly spread large number of sensor nodes along an area of interest—can be any [41], [42]. We consider a traditional distributed detection topology [37], [43], where each sensor communicates its observations (test statistic) to a central node. The central node, called also a fusion centre (FC), makes a global decision about the state of the monitored area based on the sensors observations [37]. There is no feedback channel from the fusion centre to the sensors. We implicitly assume availability of a wireless communication between each sensor and the central fusion node in the network. However, we do not model the communication part of the network since it is not within the scope of this work. Our focus is on making statistical inference about monitored phenomena.

Further, we assume that the decision task at each sensor is formulated as a binary hypothesis test. Each sensor observes its environment by sampling a phenomenon of interest. The state of the phenomenon is represented by the empirical cumulative distribution function (EDF) constructed from the samples. We relax the stringent requirement that the EDF is a priori known at each node. Instead, the function must be learnt by each sensor before the beginning of its operational regime. In effect, it might be unrealistic to assume prior knowledge of EDF in IoT scenarios with a large number (of the order of tens of thousands to hundreds of thousands [41], [42]) of sensors.

The main assumption we make is that the observations of each sensor are independent from those of the remaining nodes. This hypothesis seems realistic for a scenario where each sensor is responsible for monitoring a different phenomenon (target) within a common for all nodes field. Another case that might comply with this hypothesis is sensors observing the same phenomenon, such as temperature, but in different spatial locations. The independence assumption might be a plausible one as well for a large-scale sensor network, which covers an extended spatial area, where different phenomena can occur in separate locations of the monitored field rather than one event being spread along the entire region. The task of the fusion centre, under these settings, might be to decide if the network is under nominal conditions or if those have changed. Another relevant task is to detect the location or area where the change in nominal conditions occurs.

1.2 Goal

The main problem addressed in the thesis is performing statistical inference in Internet of Things scenarios, which typically feature a massive number of sensor nodes. Since prior knowledge of probability models might not be available in multinode networks in general, the goal is to propose a statistical inference framework based on a non-parametric detection. To this end, the local detector employed at each sensor must learn the nominal conditions instead of being supplied with a probability model of an observed phenomenon. Furthermore, since we consider a traditional distributed detection topology that consists of sensor nodes and a central node (FC), a method for simultaneously evaluating the sensor test statistics is needed in the fusion centre.

1.3 Contribution

A detector [40] is implemented in each sensor node for detecting a change in nominal conditions. The detector models the observed phenomena by constructing corresponding EDFs. It applies bootstrap principles [11] and the non-parametric two-sample Anderson-Darling test [1] for calculating relevant statistics. These statistics are not used at the sensor for making local decisions but instead are communicated to a central node (the fusion centre) for making a global decision. The fusion centre employs the False Discovery Rate (FDR) control procedure [1] to simultaneously evaluate the massive number of received statistics and to control error rates.

The primary contribution of the proposed statistical inference framework is that it is tailored for large-scale multisensor networks by combining non-parametric local detection approach with multiple hypotheses testing procedure that controls error rates. The performance evaluation results bring insights and novel ideas for improving its statistical power.

Thesis organisation. The reminder of the thesis comprises a brief look into the theoretical foundations behind the local detector in Chapter 2, multiple hypotheses testing in Chapter 3 as well as the false discovery rate (FDR) control procedure in Chapter 4. We continue in Chapter 5 with the experimental design for the simulations reported in Chapter 6, where we focus on the impact of different parameters on FDR. Next, in Chapter 7 we evaluate the performance of the proposed large-scale statistical inference approach under diverse conditions. In Chapter 8 we review and discuss prior art that explores FDR in the context of wireless sensor networks. We draw conclusions and outline potential next research directions in Chapter 9.

Chapter 2

Local detector

A local detector [40] is employed at each sensor node in the network. The key idea is that the nominal conditions of the observed phenomenon do not need to be known a priori, but instead are learnt from data. The decision problem at the detector is formulated as a binary hypothesis (2.1.1)—the goal is to detect a departure from nominal conditions when such occurs. In particular, the problem is defined as a two-sample test, which is evaluated via the Anderson-Darling test statistic (2.1.2). The nominal conditions observed by each sensor are expressed through a probability distribution, which is approximated by bootstrapping (2.1.3).

2.1 Building blocks

2.1.1 Binary hypothesis testing

Binary hypothesis testing is a common statistical problem in which the null hypothesis $\mathcal{H}_0$ is well-defined and often stated as 'the same' or 'no difference', whereas the alternative, non-null, hypothesis $\mathcal{H}_1$ is its counterpoint: 'not the same' or 'different'.

One typical binary hypothesis test is that of detecting a shift in the mean of a normal distribution. It is formulated as follows:

$$\mathcal{H}_0 \quad : x_1, \ldots, x_m \quad \overset{iid}{\sim} \quad \mathcal{N}(\mu_0, \sigma^2) \tag{2.1}$$

$$\mathcal{H}_1 \quad : x_1, \ldots, x_m \quad \overset{iid}{\sim} \quad \mathcal{N}(\mu_1, \sigma^2). \tag{2.2}$$

The nominal conditions (those under the $\mathcal{H}_0$) are assumed known. In the experimental part, Chapter 5 and Chapter 6, we model them with a zero-mean $\mu_0 = 0$ normal distribution with known variance σ^2; the alternative

hypothesis is modelled with a Gaussian with non-zero, positive mean μ_1 and the same variance σ^2 as under nominal conditions.

In chapters 5 and 6 we apply the parametric z-test to this problem. The z-test determines if a sample from a normal distribution with a known standard deviation σ has a particular (population) mean μ_0. The test statistic is as follows:

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}, \tag{2.3}$$

where $\bar{X}$ denotes the sample mean, μ_0 is the population mean (that is, mean under $\mathcal{H}_0$), σ is the common, for the population and sample, standard deviation and n is the sample size. The test statistic follows a normal distribution under the null hypothesis, and a p-value can be readily obtained (see (2.19) below).

Another common problem is that of detecting a difference in the variances of two Gaussians with equal mean μ. It is formulated as follows::

$$\mathcal{H}_0 \quad : x_1, \ldots, x_m \quad \overset{iid}{\sim} \quad \mathcal{N}(\mu, \sigma_1^2) \tag{2.4}$$

$$\mathcal{H}_1 \quad : x_1, \ldots, x_m \quad \overset{iid}{\sim} \quad \mathcal{N}(\mu, \sigma_2^2). \tag{2.5}$$

In Chapter 6, we consider zero-mean ($\mu = 0$) Gaussian distributions and apply the F-test, which has the test statistic as follows:

$$F = \frac{s_1^2}{s_2^2}, \tag{2.6}$$

where s_1^2 and s_2^2 are the variances of the two samples (i.e. sample variances). The null distribution of the test statistic F has an F-distribution.

In the local detector, the binary hypothesis testing problem is formulated in terms of probability distributions:

$$\begin{aligned} \mathcal{H}_0 : &\quad \text{data samples } \mathbf{X} \text{ and } \mathbf{Y} \text{ obey the same distribution} \\ \mathcal{H}_1 : &\quad \mathbf{X} \text{ and } \mathbf{Y} \text{ do not obey the same distribution,} \end{aligned} \tag{2.7}$$

where the data $\mathbf{X}$ is called ambient sample and $\mathbf{Y}$ observation sample [40]. The data in $\mathbf{X}$ is *reference* or *training data* as it is obtained under nominal conditions. In signal processing, for instance, nominal conditions refer to the condition when only noise but no signal is observed. The *observation sample* $\mathbf{Y}$ is recorded during the actual detection operation of the local detector and is used to test whether the nominal conditions are present or if these have changed along the observational cycle.

Error rates. Any binary hypothesis testing procedure is subject to two errors. *Type I error* occurs when the evaluated $\mathcal{H}_0$ hypothesis is a true null but it is rejected. Depending on the field of study, type I error is also called a *false alarm* or *false positive*. A measure of the error of the first kind is the false alarm probability, which we denote by p_{fa}. *Type II* error occurs whenever the test fails to reject the null hypothesis $\mathcal{H}_0$ when in fact it is false. The error of the second kind is called a *miss* or a *false negative* and is measured by the probability of missed detection, p_{md}. Classical statistical decision theory assigns a bound—a *level of significance* α—to the type I error. The objective is to minimise the type II error or equivalently maximise the *power of the test* $1 - \beta$, subject to this constraint.

Decision about $\mathcal{H}_0$	$\mathcal{H}_0$	
	True	**False**
Fail to reject	correct inference $1 - \alpha$	*type II error* β
Reject	*type I error* α	correct inference $1 - \beta$

Table 2.1: Binary hypothesis testing—correct decision rates and error rates.

2.1.2 Anderson-Darling test statistic

The Anderson-Darling test [1], [2] belongs to the goodness-of-fit tests that compare the empirical distribution function (EDF) of a data sample with a specified theoretical distribution. The EDF F_n of an i.i.d. data sample $\mathbf{X} = \{X_1, ..., X_n\}$ of size n can be expressed by:

$$F_n = \begin{cases} 0 & \text{if} \quad x < X_{(1)} \\ i/n & \text{if} \quad X_{(i)} \leq x < X_{(i+1)}, \quad i = 1, ..., n-1 \\ 1 & \text{if} \quad X_{(n)} < x, \end{cases} \tag{2.8}$$

where $X_{(i)}$ are the rank ordered data points of the sample $\mathbf{X}$: $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$. Clearly, the empirical function $F_n(x)$ is the proportion of the X_i, $i = 1, ..., n$ that are less than x:

$$F_n = \frac{\#\{i : X_i \leq x\}}{n}. \tag{2.9}$$

In other words, the empirical function puts an equal mass to each observation point.

The Anderson-Darling (AD) test is a modification of the widely used Kolmogorov-Smirnov (KS) test, which we briefly recall first. The KS test is a non-parametric hypothesis test that quantifies the *largest difference* between the empirical F_n of an i.i.d. data sample $\mathbf{X}$ and a particular distribution function F.

The Kolmogorov-Smirnov *goodness-of-fit test* is formulated as:

$$\begin{aligned} \mathcal{H}_0: &\quad \text{data sample } \mathbf{X} \text{ follows a specified distribution } F \\ \mathcal{H}_1: &\quad \mathbf{X} \text{ does not follow the specified distribution } F \end{aligned} \tag{2.10}$$

and the KS test statistic that is based on the largest vertical discrepancy between the cumulative F and empirical F_n distribution functions is:

$$\tau_{\mathrm{KS}} = \sup_x |F(x) - F_n(x)| \,. \tag{2.11}$$

The test statistic does not depend on the underlying distribution function F. One of its known disadvantages is that it is more sensitive to the centre of the distribution rather than its tails, and in (2.10) form, the cumulative distribution function (CDF) F must be fully specified. The distributions are assumed to be continuous.

The *two-sample* version of the KS test reads as follows:

$$\begin{aligned} \mathcal{H}_0: &\quad \text{data samples } \mathbf{X} \text{ and } \mathbf{Y} \text{ come from identical distributions, } F_n = G_m \\ \mathcal{H}_1: &\quad \mathbf{X} \text{ and } \mathbf{Y} \text{ do not come from the same distribution, } F_n \neq G_m, \end{aligned} \tag{2.12}$$

where F_n is the EDF (2.9) of $\mathbf{X}$ and G_m is the EDF of data sample $\mathbf{Y}$. The two-sample KS test statistic is given by:

$$\tau_{\mathrm{KS}_2} = \sup_x |F_n(x) - G_m(x)| \,, \tag{2.13}$$

An alternative to the two-sample KS test is the Cramér-von Mises (CvM) test based on the integral of the squared discrepancies between the two EDFs:

$$\tau_{\mathrm{CvM}_2} = \frac{nm}{n+m} \int_{-\infty}^{\infty} [F_n(x) - G_m(x)]^2 dH_{n+m}(x), \tag{2.14}$$

with the EDF of the combined X and Y sample of size $N = m + n$ given by:

$$H_{n+m} = \frac{n}{N} F_n(x) + \frac{m}{N} G_m(x). \tag{2.15}$$

The two-sample Anderson-Darling test introduces a weighting function $\psi(x)$ into (2.14), so that:

$$\tau_{\mathrm{AD}_2} = \frac{nm}{n+m} \int_{-\infty}^{\infty} \psi(x)[F_n(x) - G_m(x)]^2 dH_{n+m}(x), \tag{2.16}$$

where $\psi(x)$ counteracs the fact that the descripency between the EDFs is becoming smaller in the tails since the distribution functions approach 0 and 1 at the extreems:

$$\psi(x) = \frac{1}{H_{n+m}(x)(1 - H_{n+m}(x))}, \tag{2.17}$$

that is, the discrepancy is weighted by a factor reciprocal to its variance, and has the effect of giving greater importance to observations in the tails than do most of the EDF statistics [4]. Consequently, the AD test is expected to be a more powerful test in detecting alternative hypotheses that have high probability of generating observations in the tails [4].

For practical purposes, the AD test statistic (2.16) is expressed by [3]:

$$\tau_{\mathrm{AD}_2} = \frac{1}{mn} \sum_{i=1}^{N-1} \frac{(M_i N - ni)^2}{i(N-i)}, \tag{2.18}$$

where M_i is the number of observations from the data sample $\mathbf{X}$ less then or equal to the i-th smallest in the combined sample.

The two-sample AD test is extended in [5] to a k-sample version. It is motivated by the need to test for a differences in several independent samples.

The asymptotic distribution of the AD test statistic is discussed and some critical values are tabulated in [1], [3] and [5]. The distribution of the test statistic under the null hypothesis can be constructed by recording the distribution for all $N!/n!m!$ in the case of two samples [3] or rank permutations in the case of a k-sample test [5]. However, as k and the total size of the pooled sample N (for τ_{AD_2}, $N = n + m$) grows, the computations and tabulations increase sharply due to the enourmous number of permuations required. Furthermore, Pettit [3] and Scholz and Stephens [5] show that the distribution of τ_{AD} converges to a limiting distribution.

To overcome such computational and tabulation effort, which might be impractical in IoT scenarios, the essencial ideas of the bootstrap method are employed in the local detector as explained below.

2.1.3 The bootstrap method

The bootstrap is a technique commonly used for estimating the distribution of a parameter θ of interest when its accuracy is not known and can not

be calculated analytically. The general set-up is that there is a single data sample $\mathbf{X} = \{X_1, ..., X_n\}$ of length n from a distribution F, $P(X_i \leq x) = F(x)$, and a parameter θ of F must be estimated. There is also an estimator $\hat{\theta}$, which depends on the data sample and which can be used to estimate θ from data. The distribution of the estimator can be obtained by repeating the data-generation experiment a large number N of times and calculating $\hat{\theta}$ for each of the N data sets of length n. However, collecting new data sets might not be feasible nor efficient. The bootstrap method overcomes this problem by approximating the distribution F of the data-generating process and using this approximated $\hat{F}$ distribution instead of the true F. The non-parametric bootstrap, **Algorithm** 1, is used when the distribution function F is not known but its EDF F_n is learnt from data.

Algorithm 1 The non-parametric bootstrap algorithm [11]

Step 1. Obtain a data sample $\mathbf{X} = \{X_1, ..., X_n\}$
for $b = 1, \ldots, N$ **do**
  Step 2. Sample a new data set $\mathbf{X}^*$ of size n from $\mathbf{X}$ with replacement.
  Step 3. Estimate θ from $\mathbf{X}^*$. Denote the estimate by $\hat{\theta}_b$.
end for
Step 4. Consider the EDF of $\{\hat{\theta}_1^*, ..., \hat{\theta}_N^*\}$ as an approximation to the true distribution of $\hat{\theta}$.

The local detector estimates the distribution of the test statistic τ_{AD_2} (2.16) using resampling similar to Algorithm 1 on which we elaborated below.

2.2 Algorithm

We adapt the local detector proposed in [40] to the settings of the Internet of Things scenario, where a massive number of sensors observe an area/target(s) and the overall decision about the entire field of observation is made in a central node. The operational mode of the local detector [40] comprises a *training*, *observation* and *decision* phases. The modification we make is in the final phase, which we call a *hypothesis testing*. It involves the calculation of a test statistic and determination of its corresponding p-value. In contrast to [40], where each sensor makes a local decision, the p-value is sent to the fusion centre of the sensor network for centralised decision making.

During the **training**, the empirical distribution used to approximate the true distribution of the AD test statistic (2.18) is obtained as follows. The sensor node first records a training data sample I under nominal conditions;

that is, it samples from the unknown distribution under the null hypothesis $\mathcal{H}_0$. Then, it resamples from I a large amount B of samples $\mathbf{Z}^{\mathbf{b}*}$ each of length $N = n + m$ ($N << $ length of I). Each sample $\mathbf{Z}^{\mathbf{b}*}$ is split into two samples $\mathbf{X}^{\mathbf{b}*}$ and $\mathbf{Y}^{\mathbf{b}*}$ of length n and m, respectively. Both $\mathbf{X}^{\mathbf{b}*}$ and $\mathbf{Y}^{\mathbf{b}*}$ follow the same distribution F under $\mathcal{H}_0$ as they are sampled under the same conditions, when the null is true. Bootstrapping when $\mathcal{H}_0$ is true fulfills guideline 1 from [6]. These two samples are used to compute $\tau_{AD_2}^{b*}$ (2.18). The EDF (2.9) of τ_{AD_2} is constructed based on the calculated test statistics $\{\tau_{AD_2}^{1*}, \ldots, \tau_{AD_2}^{B*}\}$. This empirical function is used later on during the hypothesis testing phase. The training is conducted only once as long as the nominal conditions do not change along the lifetime of the sensor.

During the actual detection task only the observation and hypothesis testing phases are executed. The **observation** phase consists of recording an observational sample $\mathbf{Y}$ of length m. It is used to test whether the nominal conditions (the distribution under $\mathcal{H}_0$) have changed. The **hypothesis testing** phase consists of randomly sampling data $\mathbf{X}$ of length n from the training data I. Then, the test statistic τ_{AD_2} is computed according to (2.18). The p-value of the test statistic is determined based on the calculated τ_{AD_2} and the empirical distribution of the test statistic.

We recall the definition of the p-value:

$$p = \int_{X_i}^{\infty} f_0(x)\, dx = 1 - F_0(x), \tag{2.19}$$

where X_i is the i-th observation for which a p-value is calculated, $f_0(x)$ is the probability density function of the observations under the null hypothesis $\mathcal{H}_0$, and F_0 is their cumulative distribution function under $\mathcal{H}_0$. When the p-value is determined from an EDF, it is calculated likewise:

$$p = \frac{r + 1}{B + 1}, \tag{2.20}$$

where r is the number of observations from the EDF greater than or equal to the recorded observation and B is the total number of observations in the EDF.

The local detector is summarised in **Algorithm** 2.

It is shown in [40] (see *Proposition* 1 therein) that the bootstrap detection method, on which **Algorithm** 2 is based, is valid and consistent.

Algorithm 2 Local Detector based on [40]

Training

Step 1. Obtain a training sample $\mathbf{I} = \{I_1, \ldots, I_s\}$ of length s from a data-generation process with an (unknown) distribution function F.
Step 2. Resample a large amount of samples Z^{*b}, $b = 1, \ldots, B$ of length $N = n + m$, from $\mathbf{I}$.
for $b = 1, \ldots, B$ **do**
Step 3. Split $\mathbf{Z}^{\mathbf{b}*}$ into $\mathbf{X}^{\mathbf{b}*}$ of length n and $\mathbf{Y}^{\mathbf{b}*}$ of length m.
Step 4. Calculate the test statistic τ^{b*} (2.18) using the bootstrap samples $\mathbf{X}^{\mathbf{b}*}$ and $\mathbf{Y}^{\mathbf{b}*}$.
end for
Step 5. Approximate the distribution of the test statistic τ through its empirical distribution function (2.9) constructed from $\{\tau^{1*}, \ldots, \tau^{B*}\}$.

Observation

Step 6. Obtain an observation sample $\mathbf{Y} = \{Y_1, \ldots, Y_m\}$ of length m from a data generation process with unknown distribution function G.

Hypothesis Testing

Step 7. Sample $\mathbf{X} = \{X_1, \ldots, X_n\}$ of length $n << s$ from the training data $\mathbf{I}$.
Step 8. Compute test statistic τ (2.18) for $\mathbf{X}$ and $\mathbf{Y}$.
Step 9. Determine the p-value (2.20) of the test statistic τ based on the EDF obtained during the training phase.

Communication

Step 10. Send the calculated p-value to the fusion centre.

Chapter 3

Multiple hypotheses testing

3.1 The multiplicity problem

The general problem in multiple hypotheses testing is how to simultaneously test more than one hypotheses. This is the problem we encounter in the IoT scenarios studied in this work—several sensors concurrently test for departure from nominal conditions. The easiest approach would be to test the m hypotheses separately ignoring the multiplicity. However, such a solution leads to a very high probability of false alarms as shown below.

Assume that each individual hypothesis from a total of m is tested at a significance level α. In other words, let the probability of a false alarm for each test be $p_{fa} \leq \alpha$. Further, let the m hypotheses be independent. In mathematical terms, the probability of at least 1 significant result (discovery, which we denote by d) when all null hypotheses are true is: $P(d \geq 1) = 1 - (1-\alpha)^m$. For a scenario with the modest number of $m = 100$ hypotheses, all of them being true nulls, and significance level $\alpha = 0.05$, there will be almost certainly at least one false discovery, $P(d \geq 1) \approx 0.994$, see Fig. 3.1. In fact, there can be on average 5 tests that incorrectly reject the null hypothesis since $\mathbb{E}[V] \leq \alpha m$, where V is the number of false discoveries (type I errors). Moreover, under this setting, the probability of at least one false alarm among all tests increases with the total number m of tests. In a large-scale data sets [12] with $m = 10,000$ hypotheses for instance, the number of falsely rejected null hypotheses will be 500 on average. In a single simultaneous test of the m hypotheses, the false alarms can easily surpass 500.

The problem of increased type I errors when simultaneously testing multiple hypotheses is known as *the multiplicity problem*.

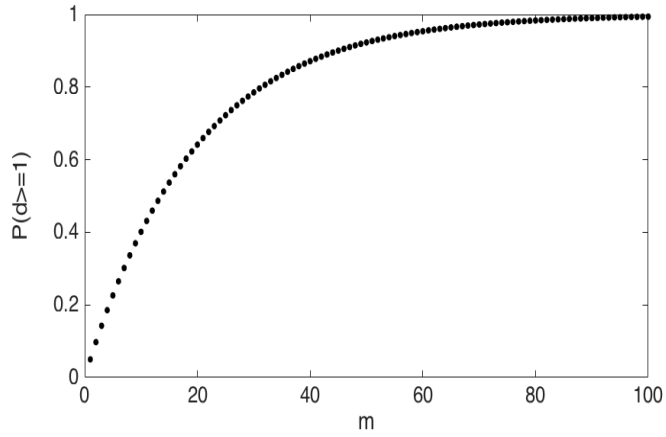

Figure 3.1: Probability of at least one false alarm $P\,(\,d \geq 1\,)$ among a total number of m independent tests when each of the m hypotheses is evaluated individually at a conventional significance level $\alpha = 0.05$.

3.2 Error rates and control procedures

In the multiple testing task there is a total of m hypotheses simultaneously tested. The number m_0 of true null hypotheses is unknown, $0 \leq m_0 \leq m$, and so is the number of false hypotheses $m_1 = m - m_0$. The proportion of true null hypotheses, $\pi_0 = m_0/m$, is not a directly observable variable either.

The contingency Table 3.1 lists the number of errors that can occur in multiple hypotheses testing.

	Declared non-significant	Declared significant	Total
True null	U	V	m_0
Non-true null	T	S	$m - m_0$
Total	$m - R$	R	m

Table 3.1: Number of correct and erroneous decisions when testing m null hypotheses of which m_0 are true.

The total number m of tested hypotheses and the total number R of rejected hypotheses are known variables, but none of the remaining variables in Table 3.1 can be observed, namely the number V of type I errors, the number T of type II errors, the correctly non-rejected null hypotheses U as well as the number S of correct detections are unobservable random variables.

Similar to the binary hypothesis set-up, in multiple comparison testing, the objective is to make as many true discoveries as possible while keeping the number of type I errors as low as possible. In contrast to the binary hypothesis testing however, the error measures are defined differently. They reflect the two general approaches to the multiplicity problem.

The traditional approach replaces the individual test requirement of $p_{fa} \leq \alpha$ with a new one. It applies a threshold α on the global probability of one or more false rejections among the m hypotheses. Hence, it controls the number of false discoveries V. Since the collection of hypotheses being tested simultaneously is called a "family", the latter probability is called a family-wise error rate (FWER). It is given by:

$$\text{FWER} = P\,(\,V > 0\,). \tag{3.1}$$

and the new test requirement is:

$$\text{FWER} \leq \alpha. \tag{3.2}$$

Strong control of the type I error rate refers to control of the FWER for all possible constellations of true and false null hypotheses [8]; that is, for any subset of $\mathcal{H}_0$ hypotheses. On the contrary, weak control of the FWER means that (3.2) holds only when all null hypotheses are true.

The more recent approach to the multiplicity problem controls a false discovery rate (FDR). This rate is defined through the false discovery proportion (FDP) [7], which is a random variable expressed by the number of true discoveries S and the number of false discoveries V:

$$Q = \frac{V}{S+V} = \frac{V}{R}, \qquad 0 \leq Q \leq 1. \tag{3.3}$$

By convention, when the total number of discoveries $R = S + V$ equals 0, also the FDP $Q = 0$ as no error can occur. Q is unknown random variable for the reasons mentioned earlier. The false discovery rate is the expectation of the false discovery proportion [7]:

$$Q_e = \mathbb{E}\,[\,Q\,] = \mathbb{E}\left[\frac{V}{R}\right]. \tag{3.4}$$

The two error rates are related through the following inequality [8]:

$$\mathbb{E}\,[\,Q\,] = Q_e \;\leq\; P\,(\,Q > 0\,) = P\,(\,V > 0\,) = \text{FWER}, \tag{3.5}$$

which implies that the FWER-type procedures control the FDR too. Also, FDR-type procedures control the FWE when $\pi_0 = 1$; that is, FDR controls FWE in the weak sense [7]. Further, FDR $<$ FWER implies a higher power of the FDR control procedures whenever there are some false null hypotheses.

Other type I error measures defined in the multiple hypotheses testing literature (for a summary see for instance [10]) are: per-comparison error rate (PCER), which is the expected value of the number of type I errors V versus the total number of hypotheses m, PCER=E$[V]/m$, and the per-family error rate (PFER), which is the expected number of type I errors, E$[V]$.

The FWER and FDR control procedures are step-wise procedures that are stated in terms of the p-values of the tests. The basic FWER control procedure—the Bonferroni method—is a single-step procedure. It rejects each hypothesis at a common cut-off value α/m [8]. In particular, the family-wise error rate is controlled at level α, FWER $< \alpha$, by individually testing each hypothesis $\mathcal{H}_i$ from a total of m hypotheses at α/m. Stated another way, the procedure rejects the i-th hypothesis, $\mathcal{H}_i$, if $p_i \leq \alpha/m$. For large-scale hypotheses testing problems, the power of the Bonferroni procedure, namely the ability to detect cases in which some of the null hypotheses are false or equivalently some of the alternatives are true, is practically 0. This can be explained by the fact that control of FWER requires each of the individual hypotheses to be tested at a much lower level then α when their total number m is large.

The multiplicity problem discussed at the beginning—the increased number of type I errors when simultaneously testing a family of hypotheses—can be addressed by methods that control any of the discussed error rates. Nevertheless, their objectives and accordingly application areas differ. FWER is relevant when the overall conclusion is likely erroneous if even one of the true null hypothesis $\mathcal{H}_0$ is falsely rejected. The FDR control porcedure is of interest in those cases for which the overall decision is not necessarily erroneous even if some of the $\mathcal{H}_0$ hypotheses are falsely rejected. Furthermore, the FDR control assumes that when many of the tested hypotheses are rejected, it may be preferable to control the proportion of errors rather than the probability of making any error [18]. In other words, the FDR control procedure is more liberal in identifying discoveries than are FWER-based algorithms, at the cost of increased number of type I errors [18]. Hence, FDR finds application whenever the interest is in making discoveries.

Chapter 4

False Discovery Rate Control

The main advantage of the FDR procedure is the gain in power compared to the FWER control procedures (Chapter 3). As pointed out in [18], when the number of hypotheses m reaches hundreds, "addressing the multiplicity problem by controlling the FWE is overwhelmingly conservative". In IoT applications, similar to microarrays, this number m can be well above the hundreds. This motivates our choice of the FDR control procedure for simultaneously evaluating multiple hypotheses in IoT settings. We provide a brief overview of the method below.

4.1 The procedure

The Benjamini-Hochberg approach to the multiplicity problem can be seen as a control of the false discovery rate subject to maximasing the total number of discoveries [7]. The procedure is summarised in **Algorithm** 3. It guarantees that the average false discovery proportion Q_e (in other words, the false discovery rate) is less than a predefined threshold q^* ($Q_e \leq q^*$), when the m test statistics are independent, see *Theorem 1* in [7]. The same result is proven to be valid as well under certain positive correlation structures [18].

Algorithm 3 The FDR algorithm of Benjamini and Hochberg [7]

Step 1. Obtain the p-value of the test statistic of each hypothesis:

$$p_i := P\,(X \geq x_i), \quad i = 1, \ldots, m.$$

Step 2. Rank order the calculated p-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$, so that $p_{(1)}$ represents the most extreme tail probability and $p_{(m)}$ is the least extreme tail probability.

Denote by $H_{(i)}$ the null hypothesis corresponding to $p_{(i)}$.

Step 3. Find the largest i for which

$$p_{(i)} \;\leq\; \frac{i}{m} q^* \tag{4.1}$$

holds. Let this be the kth p-value in the ordered list if such a p-value exists.

Step 4. Reject all null hypotheses, which correspond to the first $\boldsymbol{k}$ p-values from the ordered list; that is, reject $H_{(i)}$ with $i = 1, 2, \ldots, k$.

Further, according to the *Lemma* in [7], for any independent p-values corresponding to the m_0 true null hypotheses, and for any values that the m_1 p-values corresponding to the false null hypotheses can take, the procedure defined by **Algorithm** 3 satisfies the inequality:

$$Q_e \;= \mathbb{E}\,[\,Q\,] \;\leq\; \frac{m_0}{m}\, q^* \;=\; \pi_0\, q^* \;\leq\; q^*. \tag{4.2}$$

Hence, (4.2) suggests that the FDR control procedure is characterised by an increase in the false discovery rate Q_e when the proportion of true null hypotheses π_0 grows. This could be intuitively explained by the fact that when π_0 grows, the probability that the value of the test statistic of some of the true hypotheses will belong to the tail of the distribution under the null hypothesis also increases. In other words, the phenomenon described at the beginning as the multiple testing problem[1] is still present and observed.

In summary, the FDR procedure guarantees that in the long run, the proportion of false positives (type I errors) will be below q^*. The particular value of the actual FDR, however, depends on the true nulls fraction $\pi_0 = m_0/m$ as implied by (4.2). When the fraction of true null hypotheses is (very) small, the mean FDP will also be (much) smaller than the threshold q^*.

[1]The larger is the number of true null hypotheses tested, the higher is the probability that some of them will be erroneously rejected.

When all of the null hypotheses are true, the FDR procedure controls the family wise error rate as mentioned earlier. This explains the usual choice of the FDR threshold q^* at the conventional levels for α [18]—the probability of making at least one type I error. Otherwise, when the number m_1 of alternative hypotheses increases (correspondingly π_0 decreases), the number of correctly rejected by the FDR procedure hypotheses S tends to be larger [7]. Consequently, the FDR power increases too and is larger than the FWER power in general.

4.2 Discussion

The FDR control procedure has found application primarily in biomedical studies and most recently in microarray analysis[2]. It was in the past few years when FDR was examined in the context of wireless sensor networks for the frst time (see Chapter 8). Therefore, we briefly pause to discuss the main conceptual difference between the FDR application to these two different research areas.

In microarray analysis, the usual working assumption is that the genes of a group of patients differ from the genes of a group of healthy individuals (the control group). The goal is to detect as much as possible "significant" genes that are responsible for the examined "disease". In wireless sensor networks the goal is to detect if a change in the observed phenomenon has occurred. For some applications, in addition to detecting occurrence of a phenomenon, it might be relevant to know the exact spacial location of the observation. However, whereas the "significant" genes are directly identified by the FDR control procedure (subject to some error rate), in wireless sensor networks the identification of the sensors with "discoveries" is not intrinsic. In fact, in networks with a massive number of sensors and the common in the scientific literature random sensor distribution set-up [41], it is clearly unrealistic to assume that sensors locations are implicitly known (see Section 6.1 in [42] for instance). Phenomena location estimation requires additional mechanisms to be employed too.

4.2.1 Original FDR control

In medical research Benjamini–Hochberg FDR algorithm has been applied to make inference about the individual hypotheses and to decide, which of them are false; that is, in making discoveries [17].

[2]The DNA microarrays is a technology, which allows for collecting quantitative measurements for the expression of thousands of genes [12], [13].

Goeman and Solari [9] draw attention to two aspects of the Benjamini–Hochberg algorithm [7] with regard to its use.

Firstly, the FDR control holds only for the full set of rejected hypotheses, but not for the individual hypotheses. This is in contrast to FWER control, which guarantees that if the family wise error rate for a rejected set is below level α, then every hypothesis $\mathcal{H}$ from the rejected set $\mathcal{R}$ is a type I error with probability less than α. In other words, the FDR algorithm implies error control on average, over all hypotheses $\mathcal{H} \in \mathcal{R}$, but does not guarantee that the probability of each individual hypothesis being a false positive is less than α.

Secondly, FDR control at level α only controls the false discovery proportion Q in expectation, but the actual proportion of false discoveries in the rejected set can be substantially larger (or smaller) than α according to a large-scale genomic study (see [9] and reference [65] therein). The variability of the false discovery proportion Q is due to the rejected set being a random variable and according to [9] (and reference [24] in there) FDP is especially variable when the p-values are dependent.

Also, the FDR control mechanism is compared to (empirical) Bayesian FDP estimation [9]. In FDR control, if the experiment is repeated many times, the FDP on average will be less than α. The rejected set is a random variable. In FDP estimation, the rejected set is fixed and an estimate of the confidence interval is obtained, which is the random variable in this case. For a large number of confidence intervals, the true FDP is contained in the confidence interval at least $(1 - \alpha)$ proportion of the time.

4.2.2 Empirical Bayes FDR approach

We studied the applicability of the Efron's two-group model [12] to the Internet of Things setting. Before discussing its suitability for distributed detection, however, we first summarise the approach.

The two-group model [12] assumes that the m tests are true with probability π_0 and false with the complementary probability $(1-\pi_0)$. It is assumed that the prior probability of the proportion of true hypotheses, π_0, is at least 0.9; other assumptions are discussed below. Efron's approach provides point estimates for FDP of arbitrary sets and for individual hypothesis, assuming that the test statistics corresponding to the hypotheses are drawn from a mixture distribution:

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z), \tag{4.3}$$

with z-values having density either f_0 if null and f_1 if non-null. In contrast to the original FDR control [7], Efron suggests that independence is not required

(this statement is questioned by Benjamini in [14]). Morris highlights in [16] that exchangeability[3] in the two-group model is assumed in two ways. First, π_0, the proportion of true null hypotheses should not depend on i (i is the i-th observation among a total of m observations). Second, the density f_0 under the null hypothesis $\mathcal{H}_0$ as well as the density f_1 under the alternative hypothesis $\mathcal{H}_1$ must be the same for all observations $i = 1, 2, ..., m$.

In comparing FDR and local FDR[4], Cai [15] argues that the local false discovery rate is more meaningful quantity to be controlled by the multi-testing procedure than the p-value as it unifies both global error control and individual case interpretation (thus overcoming the limitation of FDR control discussed in [9], which is the first one listed above). In particular, Cai suggests using FDR to select a list of non-null candidates and use the local fdr to differentiate the level of certainty in the list. However, Cai [15] admits that the optimal estimation of the three fundamental quantities of the two-group model of Efron is a challenging problem.

4.2.2.1 IoT perspective

In addition to the remarks made in the studies that focus on large-scale genomic data, we make few more observations from the perspective of distributed detection in IoT scenarios.

In the two-group model of Efron, the $f_0(z)$ is the distribution of the z-scores under the null hypothesis. The z-scores are obtained from the p-values under the null hypothesis; that is, the p-value is transformed into a z-score with the inverse of the normal distribution. Since the p-values under $\mathcal{H}_0$ follow a uniform distribution $U[0, 1]$, the mapping of a p-value under $\mathcal{H}_0$ by the inverse normal distribution yields a z-score that obeys the standard normal distribution: $z = \Phi^{-1}(p) \sim \mathcal{N}(0, 1)$ [38], where Φ is the standard normal CDF. This holds for a single test as long as the test is valid, namely the assumptions about the test under $\mathcal{H}_0$ do hold. Otherwise, the theoretical null might not be $\mathcal{N}(0, 1)$. Since we are simultaneously considering a massive number of hypotheses, the distribution of z-scores is generated by correspondingly high number of p-values. The theoretical distribution is $\mathcal{N}(0, 1)$ if all the tests are valid and there is not too strong correlation between the

[3]Exchangeability refers to the the joint probability distribution of a set of variables remaining unchanged under arbitrary permutations of the indices of the variables [36].

[4]Local fdr is defined as the posterior probability of $\mathcal{H}_0$ [16]:

$$fdr(z) = P(\mathcal{H}_0|Z = z) = \frac{\pi_0 * f_0(z)}{\pi_0 f_0(z) + \pi_1 f_1(z)}. \tag{4.4}$$

tests. However, in practical scenarios, where existance of correlations might not be feasible to check, the assumption of normality might be too strong. The null distribution affects directly the final results as even a small divergence of the adopted from the true null distribution can yield very different testing results [12]. Therefore, the exact knowledge or correct estimation of the true null distribution seems fundamental. However, its estimation is an admittedly complex problem [12], [13], [15] even in the microarray case, where all the data is readily available and processing is done without (any comparable to IoT) time-constraints. Due to the typical for wireless sensor networks transmission and energy limitations, the estimation of the empirical null seems even more challenging in the distributed scenario. Commonly, the test results, but not all the measurements, are transmitted to the FC.

The empirical Bayes approach has been developed in response to the need of statistically analysing data originating from genomic studies. Although according to Efron the applicability of the approach is not constrained to the microarray alone [13], it does have specific features:

- central versus distributed availability of statistical data

  All data (gene expression levels) collected with the microarray technology is at the disposal of the statistician, which is in sharp contrast with the availability of data in the distributed detection scenario considered in this work. Any data (test statistics) must be sent to the fusion centre of wireless sensor networks under transmission and energy limitations.

- delay-tolerant vs delay-intolerant application

  Data analysis in the microarray does not have the stringent time constraints typical for IoT, where the detection of an event/decision making must be performed as fast as possible.

- automated processing vs human intervention

  The detection of an event at the fusion centre based on the received test statistics is usually done in a delay-intolerant manner and therefore follows a decision rule implemented as an algorithm rather than relaying on a human intervention as is the case in microarray analysis, where the statistician (its interpretation of the results, (pre)selection of a set of genes/hypotheses) plays a crucial role.

In summary, the empirical Bayes approach raises few questions about its applicability to IoT area, which can be reduced to the validity of the hypothesis that the null distribution of the test statistic follows a normal distribution.

Chapter 5

Experimental design

In the remaining part of the thesis we examine the performance of the proposed inference framework. Before delving into the simulation results, we describe the experimental design: the simulation framework in Section 5.1 and its validation in Section 5.1.

5.1 Simulation framework

The general framework of the simulations discussed in chapters 6 and 7 is outlined in **Algorithm 4**. The following notation is used: MC denotes the number of Monte Carlo repetitions, n the size of a sample drawn from a specific parametric distribution with parameters Θ. In the two detection examples examined in Chapter 6 for instance, the distribution is normal with user-defined value of Θ: mean μ_i and standard deviation σ_i; that is, $\mathcal{N}(\mu_i, \sigma_i^2)$. The remaining input simulation parameters are explained elsewhere in the text.

Algorithm 4 Simulation framework

1. Define: $\boldsymbol{MC}$, $\boldsymbol{m}$, $\boldsymbol{q}^*$, $\boldsymbol{\pi_0}$, Θ, $\boldsymbol{n}$
for $i = 1, \ldots, MC$ **do**
  2. Generate data and **obtain** a ***p*-value** at each sensor.
  3. Run the FDR algorithm.
  4. Collect statistics and **compute** FDP, p_d
end for
5. Calculate and **output** $\mathbb{E}[\,\mathrm{FDP}\,]$, power $= \mathbb{E}[\,p_d\,]$.

In simulation terms, π_0 corresponds to the proportion of sensors for which the data sample of size n is generated under nominal conditions. For the

scenarios in Chapter 6 for instance, the data sample n is generated from the normal distribution under the null hypothesis $\mathcal{H}_0$. For the remaining $(1-\pi_0)$ proportion of sensors the data sample is generated from the normal distribution under the alternative hypothesis $\mathcal{H}_1$. In particular, in the case of a shift in the mean, for each of the $m_0 = \pi_0\, m$ hypotheses n observations are sampled from a normal distribution with mean $\mu = \mu_0$. For the remaining $m_1 = (1-\pi_0)\, m$ local binary hypotheses the observations are generated from $\mathcal{N}(\mu_1, \sigma^2)$, where $\mu_1 \neq \mu_0$. When detecting a difference in the variance of two zero-mean Gaussians, two samples of the same size n are generated. For the true null hypotheses, $\sigma_1 = \sigma_2$, whereas for the alternative hypotheses, $\sigma_1 \neq \sigma_2$. Then, the p-values are obtained based on the generated sample(s) and by performing the corresponding (z- or F-) test at each sensor (for each local hypothesis), see **Algorithm** 5.

The FDR is calculated at the end, by averaging over the FDP obtained at each iteration:

$$Q_e = \frac{\sum_i^{MC} Q^{(i)}}{MC}, \tag{5.1}$$

where $Q^{(i)}$ denotes the FDP from the i-th simulation run, and MC is the total number of Monte Carlo realizations.

The detection probability for a given MC realisation is calculated by:

$$p_d = \frac{R-V}{(m-m_0)\vee 1} \tag{5.2}$$

$$= \frac{S}{m_1}, \tag{5.3}$$

where S is the number of non-true null hypotheses declared significant (i.e. discovered) by the FDR control procedure in a simulation run and m_1 is the number of alternative hypotheses. The number of simulated non-true null hypotheses m_1 is an input parameter, whereas S is obtained from each MC realisation. In fact, detection probability is calculated only for the simulated scenarios with $m_0 \neq m$ (that is $m_1 \neq 0$). We report the detection probability averaged over the MC simulation runs:

$$\text{power} = \frac{\mathbb{E}[\,S\,]}{m_1} \tag{5.4}$$

$$= \frac{\sum_i^{MC} p_d^{(i)}}{MC}, \tag{5.5}$$

which is the common way of computing probabilities, namely through recorded frequencies. In [10], (5.5) is denominated average power. Two other definitions of power are suggested there too [10]: probability of rejecting at least

one non-true null hypothesis $P\,(S \geq 1) \;=\; P\,(T \leq m_1 - 1)$, which is the least stringent, and the probability of rejecting all non-true alternative hypotheses $P\,(S = m_1) \;=\; P\,(T = 0)$, which is the most stringent one. We use power and average power interchanebly, and we refer to (5.5), when using any of these two terms.

Algorithm 5 describes how the p-value at each sensor is obtained for the two detection problems simulated in Chapter 6.

Algorithm 5 Generation of data and calculation of p-values

1. Difference in the means

Step 1. Generate a sample $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu_X},\, \boldsymbol{\sigma}^2)$.
Step 2. Perform a z-test to check if the mean $\boldsymbol{\mu_X}$ of the sample $\mathbf{X}$ is the same as the mean $\boldsymbol{\mu_0}$ of the population under the null hypothesis $\mathcal{H}_0$.
Step 3. Return the ***p*-value** (6) calculated by the z-test.

2. Difference in the variances

Step 1. Generate a sample $\mathbf{X}_1 \sim \mathcal{N}(0,\, \boldsymbol{\sigma_1^2})$.
Step 2. Generate another sample $\mathbf{X}_2 \sim \mathcal{N}(0,\, \boldsymbol{\sigma_2^2})$.
Step 3. Perform a two-sample F-test to check if the sample variance $\boldsymbol{s_1^2}$ of $\mathbf{X}_1$ and the sample variance $\boldsymbol{s_2^2}$ of $\mathbf{X}_2$ are equal.
Step 4. Return the ***p*-value** (6) calculated by the F-test.

In the majority of the FDR studies as well as in the related literature Chapter 8, the assumption is for independent observations. We make the same assumption here as well.

5.2 Verification

We verified the implementation of the FDR control procedure by comparing the mathematically proven result:

$$\mathbb{E}_{\text{th}}\,[\,\mathrm{Q}\,] \;\leq\; \pi_0\, q^* \tag{5.6}$$

with the simulated one $\mathbb{E}_{\text{sim}}\,[\,Q\,]$, where $\mathbb{E}\,[\,Q\,]$ denotes the false discovery rate, Q_e. The distribution of Q_e and FDR p_d that we observed with $MC = 10^4$ and $MC = 10^5$ was not in general bell-shaped. Fig. 5.1 and Fig. 5.2 show two examples: histograms and quantile-quantile plots of Q_e and the power for a scenario with $m = 100$ sensors, where a shift in the mean is tested with $MC = 10^5$ repetitions. The distributions of Q_e and power are skewed.

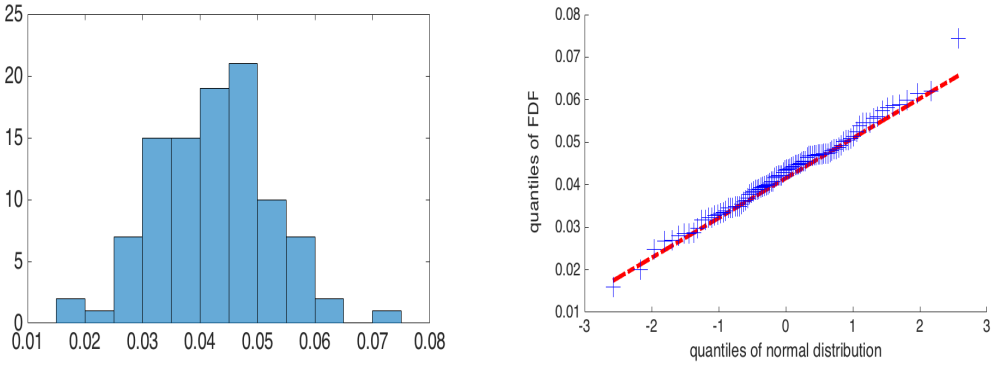

Figure 5.1: Histogram and a quantile-quantile plot against a normal distribution of the false discovery rate, Q_e.

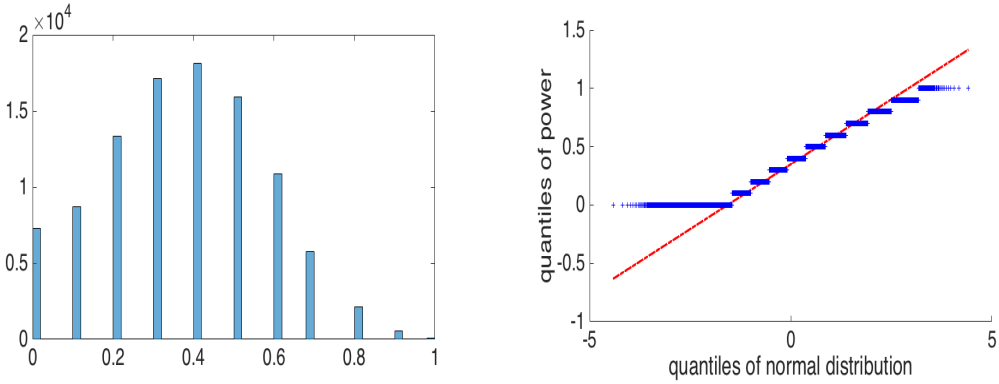

Figure 5.2: Histogram and a quantile-quantile plot against a normal distribution of the power of the FDR procedure.

Since we did not observe normally distributed FDR and power, we report bootstrapped confidence intervals (BCIs) instead of confidence intervals (CIs). The BCIs in Table A.1 are computed from $1,000$ points. Each point is obtained as an average of 100 simulated values ($MC = 10^5$ in total). The simulation results matched the analytical Q_e, see Table A.1. After verifying the simulation framework, the performance results for the scenarios of interest are obtained from $MC = 10^4$ realisations if not explicitly stated otherwise. The BCIs for the power are calculated from the 10^4 $p_d = S/m_1$ values. The reported FDR is averaged over the $MC = 10^4$ realisations.

Chapter 6

Sensitivity study of FDR control

The FDR control procedure was designed with the aim to guarantee an average false discovery proportion below a pre-defined threshold q^*. This result, $\mathrm{FDR} = \mathbb{E}[\,\mathrm{Q}\,] \leq q^*$, is analytically proven in [7]. Nevertheless, there are no similar mathematical results on the power of the FDR control. Therefore, our primary interest is in studying the FDR detection probability, whose definition is given in Section 5.1.

Although there are several studies such as [38], [10] and [39] that look into FDR power and FDR sensitivity to different input parameters, most of them are centred around the microarray experiments (as [38] and [39]) or study FDR power in comparison to other multiple hypotheses testing procedures within biological context [10].

Our objective is to study FDR first under simple IoT set-up by simulating local conditions observed by a sensor such as the signal-to-noise ratio as well as parameters that specify the state of the network as a whole such as the proportion of sensors that do not observe a change in the nominal conditions. Once we get insight into the FDR performance for a simple set-up and a large number of sensors, we proceed with examining the proposed inference framework composed by the non-parametric local detectors and FDR fusion centre in Chapter 7.

6.1 Setup

We simulate the performance of the FDR control procedure for two classical hypothesis testing problems inspired by the statistical signal processing theory [19]: detection of a signal by a shift in the mean and by a change in the variance, Fig. 6.1. The definition of the local hypothesis tests for these two problems is given in Chapter 2.

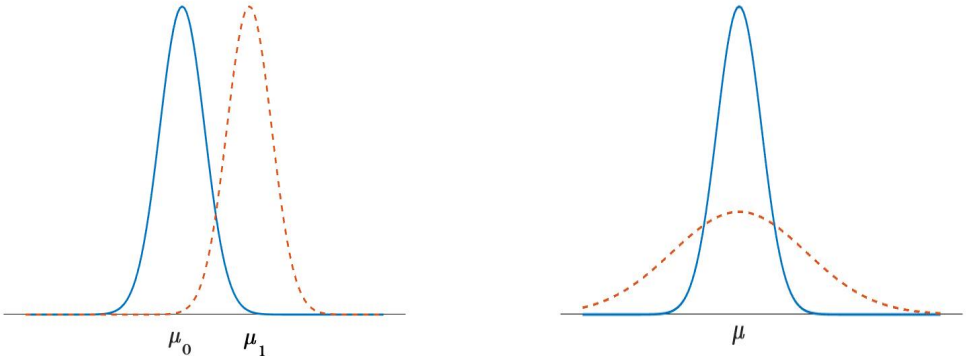

Figure 6.1: Two normal probability density functions with different mean (*left*) and different variance (*right*).

6.2 Simulation results

Sensitivity to different signal-to-noise ratio (SNR) conditions, different proportions of true null hypotheses π_0 as well as different FDR thresholds q^* is examined.

The SNR when detecting a shift in the mean is calculated by: SNR $= (\mu_0 - \mu_1)^2/\sigma^2$. When detecting a difference in the variance of two zero-mean Gaussians, the SNR is calculated by: SNR $= \sigma_1^2/\sigma_0^2$. In the signal processing literature, the signal-to-noise ratio is often expressed in decibels: $\text{SNR}_{dB} = 10\log_{10}\text{SNR}$ dB.

The choice for the two FDR bounds in the simulations is motivated by: $q^* = 0.05$ is a classical threshold in the hypothesis testing literature for the probability of a false alarm p_{fa}, whereas $q^* = 0.2$ is a common threshold in the FDR microarray literature, see for instance [12]. Another common value in biotechnological research is $q^* = 0.1$. However, to clearly observe the impact of the FDR threshold on achieved FDR power, we set $q^* = 0.2$.

The results to be reported shortly after are obtained with $MC = 10^4$ Monte Carlo realisations if not explicitly stated otherwise.

A pointer to the examined conditions and obtained results is given in Table 6.1.

6.2.1 Total number of hypotheses m

The effect of the number of hypotheses m on the power of the FDR procedure is studied for the problem of detecting a shift in the mean of a Gaussian signal. Nominal conditions (those under the null hypothesis $\mathcal{H}_0$), are modelled with

Table 6.1: Scenarios and corresponding tables with FDR results. All but Table A.2 list results for the z-test

Table	m	π_0	q^*	n	SNR(dB)
A.2	$\{10, 10^2, 10^3\}$	0.5	0.05	$\{10, 10^2\}$	-3.5
A.1	$\{10, 10^2, 10^3, 10^4\}$	0.1	0.05	$\{10, 10^2, 10^3\}$	-6
A.3	$\{10, 10^2, 10^3\}$	0.9	0.05	$\{10, 10^2, 10^3\}$	-6
A.5	10^4	0.1	$\{0.05, 0.2\}$	10^3	-30
A.4	$\{10^2, 10^3\}$	0.1	$\{0.05, 0.2\}$	10	-6
A.6	$\{10, 10^2, 10^3, 10^4\}$	0.1	0.05	10^3	$\{-6, -30\}$
6.3, 6.2	10^3	$\{0.1, 0.9\}$	$\{0.05, 0.2\}$	10^3	$\{-6, -30\}$

the standard normal distribution $\mathbf{X} \sim \mathcal{N}(0, 1)$. The conditions under the alternative hypothesis $\mathcal{H}_1$ are modelled using $\mu = 0.5$; that is, the observation sample $\mathbf{Y}$ obeys $\mathcal{N}(0.5, 1)$. As a result, $\text{SNR}_{dB} = -6$ dB. The proportion of true null hypotheses is set to $\pi_0 = 0.1$ and the FDR threshold to $q^* = 0.05$. There are $MC = 10^5$ realisations. Bootstrapped confidence intervals for FDR (Q_e) and FDR power are calculated in Table A.1, Appendix A.1.

The power of the FDR procedure decreases, under (very) few data points n, when the number m of simultaneously tested hypotheses increases. FDR power is not affected by the total number of employed sensors only when the number of observations (sample size) is sufficiently large ($n = 100$ for the studied conditions) and the SNR is relatively high. The same conclusions hold for a similar study but for $\pi_0 = 0.5$ (half of the hypotheses are true nulls) and for detecting a difference in the variance of two zero-mean Gaussians, Table A.2 in Appendix A.1.

6.2.2 Proportion of true null hypotheses π_0

We simulated the same scenario as that examined in Section 6.2.1 but for a much larger proportion of true null hypotheses, $\pi_0 = 0.9$ (only 10 % of the sensors observe a new phenomenon). The results are reported in Table A.3.

For the small sample size set-up, increase in π_0 (decrease in the number of sensors that observe departure from the nominal conditions) yields a decrease in the power of the FDR procedure. In contrast, under relatively high SNR and sufficiently large data sample ($n \approx 100$ or $n \approx 1000$ in the examined case), the proportion of true nulls π_0 does not affect FDR power.

Seen from another angle, these results indicate that when the objective is to achieve high power (close to 1), the required data sample size n increases. This conclusion bears a direct relation to a general result that under a large

number of data points the power of a detector usually increases. In summary, if the power is to be kept high or the same when the total number of sensors m or the proportion of true nulls π_0 is increased, the number of observations n that a sensor needs to collect might be large too.

6.2.3 FDR threshold q^*

To study the effect of the q^* bound on the power we considered conditions where the power is low: a small sample data set, see Table A.4, and low SNR, see Table A.5.

A larger FDR threshold improves FDR power. The penalty is an inflated number of false discoveries V. The improvement is not relevant when the SNR is low, because despite the increase in the detection probability it remains nearly zero.

The impact of q^* on FDR power is more tangible under high SNR regime (compare Table A.4 to Table A.5). Thereby, a larger threshold can be used in such cases as a remedy to low detection probability.

In addition, Table A.4 shows that for small sample size n, when the number of hypotheses is increased, the power decreases but $\mathbb{E}[V]$ increases, which is a manifestation of the multiplicity effect.

Similarly, Table A.5 shows the effect of the proportion π_0 of true null hypotheses on FDR power as well as it dependence on q^*. First, the power is larger for smaller π_0 as observed in Section 6.2.2. Second, under the same conditions, when π_0 is decreased from 0.9 to 0.1, the power is increased 10 times when $q^* = 0.05$ and about 52 times when $q^* = 0.2$. In short, the effect of π_0 on power is larger for larger q^*.

6.2.4 SNR conditions

We studied the FDR performance under high and extremely low SNR conditions, Table A.1 to Table A.3, and Table A.6. Under low SNR, the FDR power approaches 0.

6.2.5 False discovery proportion Q

According to [9], the number of cases in which the false discovery proportion exceeds the false discovery rate is often large[1]. Specifically, the users of the FDR control algorithm are warned in [9] that "control of FDR, $\mathbb{E}[Q]$, at

[1]In [9] refer to [65] therein –"FDP for a method controlling FDR at 0.10 can, for example be greater than 0.29 more than 10% of the time under independence" (page 1964).

α only controls FDP, $Q = V/(V + S)$, in expectation and that the actual proportion of false discoveries in the rejected set can often be substantially larger than α" [9] (α in [9] corresponds to the FDR threshold denoted by q^* in the thesis). This motivated us to study this question for the conditions we simulated. In particular, Table 6.2 contains the simulation parameters for which Q and Q_e listed in Table 6.3 were obtained.

Table 6.2: Power and average number V of false discoveries for the conditions studied in Table 6.3

π_0	q^*	power	mean(V)
0.1	0.05	[.99862, .99864]	4.51
	0.2	[.99987, .99988]	18.36
0.9	0.05	[.98620, .98630]	4.00
	0.2	[.99700, .99710]	22.12

Table 6.3: FDR performance results about FDP, denoted by Q, for $MC = 10^5$. The number of hypotheses tested is $m = 10^3$ and the results are from the z-test for detecting a difference in the mean between nominal conditions $\mathcal{N}(0, 1)$ and conditions under the alternative $\mathcal{N}(0.5, 1)$.

π_0	q^*	Q_e	min(Q)	max(Q)	% cases $Q \geq q^*$
0.1	0.05	[.00498, .050]	0	.0164	0
	0.2	[.01999, .020]	.0033	.0405	0
0.9	0.05	[.04480, .045]	0	.1538	33.29
	0.2	[.17990, .018]	.0291	.3506	27.88

Our simulation results show that FDP varies and can exceed q^* in certain cases. Nevertheless, it is also important to recall that the FDR procedure by design guarantees only the condition $\mathbb{E}[\,Q\,] \leq q^*$, not $Q < q^*$.

A closer look at the results in Table 6.3 suggests that the number of cases for which the false discovery proportion Q exceeds the pre-defined threshold q^* heavily depends on the proportion π_0 of true null hypotheses. The number of hypotheses and data points can further exacerbate or can diminish the observed trend.

The dependence of Q on π_0 can be explained by recalling the definition of FDP, namely $Q = V/(V + S)$, and the fact that whenever the proportion π_0 of true null hypotheses grows, so does the number V of false discoveries. The latter dependency can be intuitively understood by referring to the uniform

distribution of the p-values under the null. A p-value can land anywhere within [0,1]. In particular, by pure chance a true null hypothesis might be rejected because of small p-value. Therefore, increasing the number of true nulls can increase the number of false rejections.

6.2.6 Summary

- Increase in the number of simultaneously tested hypotheses decreases FDR power – the cost of multiplicity is manifested whenever there are only few observations per sensor and/or low SNR.
- A larger proportion π_0 of true null hypotheses decreases FDR power.
- The FDR threshold q^* controls the false discovery proportion Q in the long run. A higher threshold can improve FDR power when the power is low. This improvement depends on the particular conditions. For high SNR and small number of observations, it is tangible. For extremely low SNR such as $\text{SNR}_{dB} = -30\text{dB}$, despite some improvement, the FDR power remains close to 0.
- The false discovery proportion Q can exceed q^*. There is a clear dependence between Q and π_0: the larger the π_0 is, the larger is the number V of false discoveries and consequently the $Q = \frac{V}{V+S}$ ratio.
- The power of the FDR procedure depends on the power of the local hypothesis tests as discussed next.

6.3 How powerful is the FDR control procedure?

The Benjamini-Hochberg's false discovery rate control procedure [7] was designed for the multiple hypotheses test set-up with the aim at making as much discoveries as possible subject to a given false discovery rate (the average proportion of erroneous rejections versus all rejections). By design FDR is (much) more powerful than the FWER-type of methods (Chapter 3 and [7]), whose main drawback is, in effect, the lack of power.

Our experimental FDR performance results reported in Section 6.2 revealed cases in which the number of correct discoveries S was in the order of

at most tens when the number of true alternative hypotheses m_1 was in the order of thousands. In this section we examine what conditions yield such low power of the FDR method.

6.3.1 Dependence on the p-values

The main result—with respect to the above observation that FDR power can become arbitrarily low in specific scenarios—is that the FDR procedure is coupled with the performance of the underlying hypotheses tests that produce the p-values.

This conclusion was based on the simulation results from the scenarios discussed in Section 6.2. As an example, we observed low power of the FDR procedure for the shift in the mean set-up under SNR≈ -30dB. The null hypothesis $\mathcal{H}_0$ at each sensor is modelled with zero-mean Gaussian and the alternative hypothesis $\mathcal{H}_1$ with a positive-mean $\mu_1 = 0.5$ Gaussian. The common for both distributions variance was set to $\sigma^2 = 225$, see Table A.6. For $m = 10^3$ and $\pi_0 = 0.5$—half of the hypotheses are true nulls and the other half are true alternatives—and 10^4 Monte Carlo simulations (row 5 in Table A.6), the FDR algorithm made true discoveries S in approximately half of the simulations, namely in 4990 out of $10,000$ realisations. The maximum number of true discoveries among all the simulated realisations was $S = 19$ whereas the simulated true alternatives were $m_1 = 5,000$; the mode was at 1. For $\pi_0 = 0.9$ and $m = 10^4$ (row 8 in Table A.6), correspondingly $m_1 = 1,000$ true alternative hypotheses, the maximum number of discoveries recorded in the simulations was $S = 6$ in a single simulation run; the mode was at 1 too. In fact, the true discoveries were $S = 0$ in most of the realisations. In particular, the FDR procedure made true discoveries in about 20% of the Monte Carlo realisations. Similar results were recorder for the number of false discoveries V; that is, the realisations of V were less than 10 per experiment for both of these cases with $m = 10^3$ and $m = 10^4$, respectively. Note that our simulations confirmed the analytically proven in [7] $Q_e \leq q^*$ result[2] independent of the particular scenario. The point that these simulation results reveal is that the power of the FDR control procedure could be arbitrarily low in certain set-ups.

A look into the distribution of the p-values generated by the z-test and

[2]For high SNR regime and sufficiently large sample n for the considered shift in the mean of two Gaussians scenario, the FDR procedure correctly discovers all m_1 true alternative hypotheses (i.e. $S = m_1$) plus V additional erroneous "discoveries", $R = S + V$ discoveries in total. This additional "discoveries" V are bound by the threshold q^*; that is, FDR adheres to its design, namely $Q_e \leq \pi_0\, q^*$ (independent of the particular conditions), and under high SNR achieves high power too.

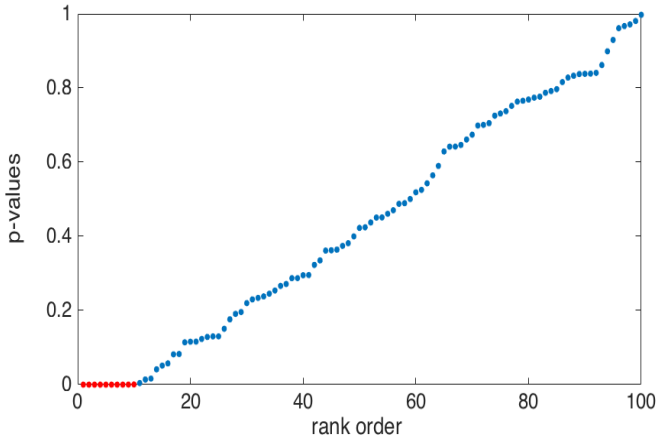

Figure 6.2: Rank ordered p-values for the scenario testing for a shift in the mean when SNR=$-$6dB, $n = 100$, $\pi_0 = 0.9$. The p-values under the alternative (in *blue*) are arbitrarily small and therefore correctly labelled by the FDR control procedure.

F-test from these simulations confirms the theoretical result that the p-values under the null hypothesis are uniformly distributed. The inspection of the p-values under the alternative hypothesis shows that this distribution determines the performance of the FDR procedure in terms of attained power.

The results in Table A.6 for high SNR= $-$6dB and large sample size $n = 10^3$ show FDR power = 1. In other words, all the alternative hypotheses are discovered by the FDR procedure independent of the other conditions: different proportions π_0 of true null hypotheses and $q^* > 0.05$ threshold[3]. All p-values under the true alternative hypothesis are very close to 0. The largest p-value among those is $\approx 10^{-37}$. As an example, Fig. 6.2 plots the same scenario but for lower number of sensors $m = 100$ (because of visualisation reasons) sample size $n = 100$, and $\pi_0 = 0.9$. All the alternative hypotheses are correctly discovered: the largest p-value under $\mathcal{H}_1$ is smaller than the critical value $p_{(i)} \leq (i/m)q^*$, namely $p_{(10)} = 0.0017 < 0.0050$. The rank ordered p-values $p_{(i)} < p_{(10)}$ are close to zero. In this particular realisation, there are no false discoveries, $V = 0$, since the samllest p-value under the true null is larger than the threshold (4.1): $p_{(11)} = 0.016 > 0.0055$.

[3]Results for $q^* = 0.2$ are not included in Table A.6 because a larger q^* always leads to a higher FDR power, as long as power is not already 1, as demonstrated in 6.2.3.

6.3.2 Dependence on π_0 and m

The simulation results listed in Table A.1, A.2 and A.3 revealed cases for which the proportion π_0 of true null hypotheses as well as the total number m of hypotheses have an impact on FDR power. Here, we look into the conditions that yield a decrease in power when π_0 and/or m are increased.

We refer again to the FDR classification threshold (4.1): $Q_e = \pi_0 q^* \leq q^*$. Let the total number of hypotheses be $m = 100$. When $\pi_0 = 0.1$, FDR control procedure will correctly discover all $m_1 = 90$ true alternative hypotheses as long as the largest p-value under $\mathcal{H}_1$ ($p_{(k)}$) satisfies $p_{(k)} \leq (90/m)q^*$. In contrast, when π_0 is increased to $\pi_0 = 0.9$, the largest p-value under $\mathcal{H}_1$ must satisfy the more stringent $p_{(k)} \leq (10/m)q^*$ condition. Consequently, the power of the FDR control procedure depends on the proportion π_0 of true null hypotheses whenever the p-values produced by the local hypothesis tests under $\mathcal{H}_1$ are not sufficiently small.

Similarly, when the total number m of hypotheses grows, the q^*/m is reduced and as a result the inequality (4.1) $p_{(i)} \leq (i/m)q^*$ becomes much more restrictive. When the p-values of the local hypotheses tests are not approximately zero, the effect of m on FDR power can be clearly observed.

Arbitrarily small p-values, equivalently power close to 1 of the local hypothesis tests, leads to high power of the FDR control procedure. It is only under such condition when FDR power does not depend on the particular network conditions given by π_0 and m.

6.3.3 Distribution of p-values under $\mathcal{H}_1$

Table A.6 (from row 5 on) shows results for very low SNR$= -30$dB. The power of the FDR procedure is less than 5% and under certain π_0 and m, it becomes 0. Fig. 6.3 shows results for $m = 1,000$, and $\pi_0 = 0.5$. The p-values are rank ordered in Fig. 6.3 under each hypotheses (and not globally as the FDR procedure does) because of visualisation purposes. It is clear that the majority of the p-values under $\mathcal{H}_1$ do not satisfy (4.1) and consequently FDR lacks power ($Pd \approx 0$).

The underlying hypotheses tests, which produce the p-values for each of the m hypotheses, determine the power of the FDR procedure[4]. Our conclusion is based on the observed distribution of p-values under the alternative hypothesis. For the z-test and the above experimental conditions—small difference in the means and high variance or good SNR but very small data

[4]The performance of the FDR procedure in terms of false discovery rate obeys $\mathrm{E}[Q] \leq q^*$ independent of the particular local hypothesis test or scenario (SNR, sample size n, proportion π_0 of true nulls).

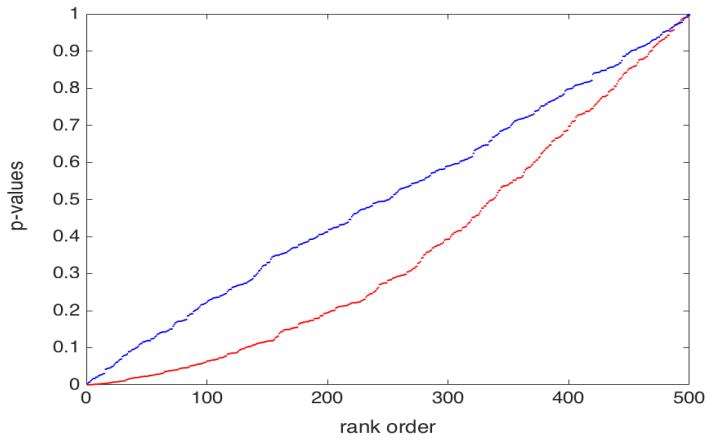

Figure 6.3: Rank ordered p-values under the null (in *blue*) and under the alternative (in *red*). A shift in the mean scenario when SNR=-30dB, $m = 1,000$, $n = 1,000$, $\pi_0 = 0.5$ and $q^* = 0.05$ is examined. The p-values under the alternative are larger than 10^{-3}, thereby the FDR control procedure cannot discover them.

sample—the distribution of the p-values under $\mathcal{H}_1$ is nearly uniform except for a set of values around 0. In contrast, the achieved FDR power is 100% when all the p-values under $\mathcal{H}_1$ are all close to zero.

The distribution of the p-values for SNR= -30dB in Table A.6 is examined in Fig. 6.4. In particular, a histogram of the distribution of p-values under true alternatives $\mathcal{H}_1$ along with a histogram of p-values under true nulls $\mathcal{H}_0$ are shown for comparison. The distribution of the p-values under the alternative can be seen as a mixture of a uniform distribution and a cluster of values near 0.

In short, some of the $\mathcal{H}_1$ p-values under low SNR regime seem to be uniformly distributed and as such are classified as true nulls by the FDR algorithm. The remaining values, which are roughly half of the total number of alternatives are clustered near 0, yet are not small enough to satisfy the FDR classification condition (4.1), Algorithm 3. Therefore, they are also incorrectly labelled as true nulls.

Another set-up for which the FDR procedure exhibited $P_d \approx 0$ was under relatively high SNR $= -6$dB but for few observation points $n = 10$, see Table A.1, A.2 and A.3. The shift in the mean case under SNR= -6dB regime and $m = 10^3$ in Table A.1 and Table A.3 is studied in Fig. 6.5, respectively. The p-values under the alternative hypothesis display similar behaviour as discussed before: grouping of p-values close to 0 and some of the p-values seem nearly uniformly distributed.

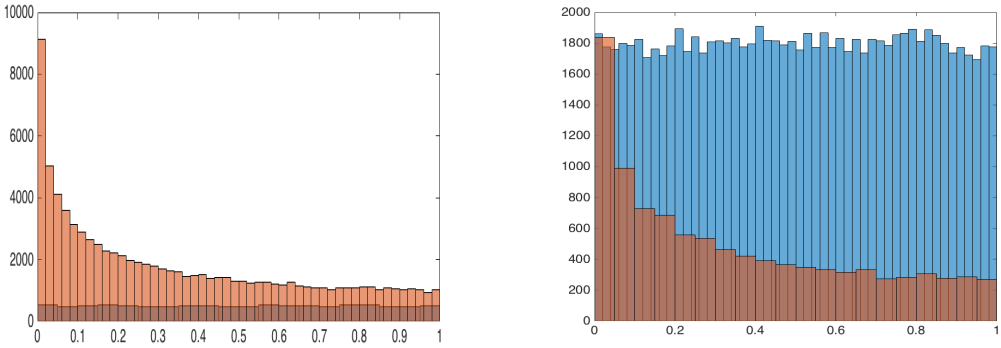

Figure 6.4: A histogram of p-values when detecting a shift in the mean of two Gaussians with the z-test for SNR=-30dB, $n = 10^4$, $m = 10^3$, (*left*) $\pi_0 = 0.1$ and (*right*) $\pi_0 = 0.9$. The distribution under $\mathcal{H}_0$ (*blue* or *dark brown* when overlapped by the distribution under $\mathcal{H}_1$ as on the left) is uniform.

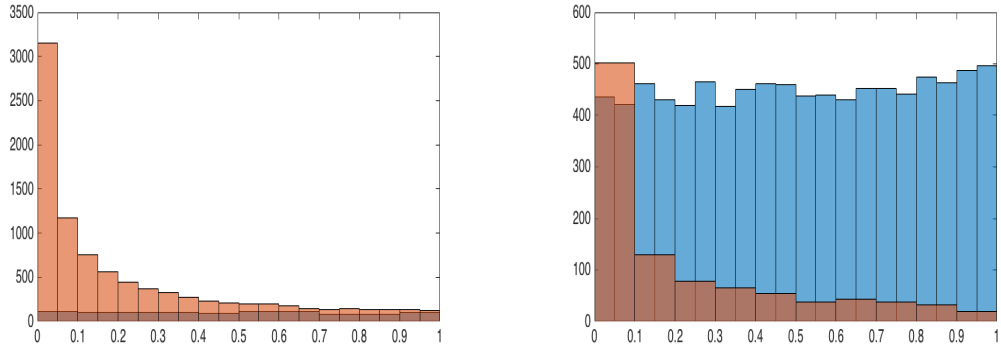

Figure 6.5: Histogram of p-values for small sample size, $n = 10$. The simulated scenario is detection of a shift in the mean of two Gaussians with the z-test when SNR= -6dB, $m = 10^3$ and $\pi_0 = 0.1$ (*left*) and $\pi_0 = 0.9$ (*right*). The p-values under the null in the right plot are seen as dark brown.

The power of the FDR procedure depends on the p-values under $\mathcal{H}_1$ and their relative position with regard to the p-values under $\mathcal{H}_0$. When the binary hypothesis test can reliably distinguish between $\mathcal{H}_0$ and $\mathcal{H}_1$ and decide on the alternative hypothesis when it is true, the test produces a very small p-value. This p-value will lead to a rejection of the null hypothesis if the test is performed only locally, at the sensor. When a multiple hypotheses test is performed globally through FDR control, this particular hypothesis will also be rejected (classified as true alternative) as long as the p-value is small and the small p-values (under $\mathcal{H}_0$ which occur by chance) do not displace

the p-values of the true alternatives "too much" to the right. In fact, FDR control procedure will correctly label all alternative hypotheses as true as long as the largest p-value among all $\mathcal{H}_1$ satisfies (4.1). For $m = 10^4$ and $q^* = 0.05$: $p_{(k)} \leq (k/2)\, 10^{-5}$, where k denotes the largest p-value.

6.3.4 Conclusions

Our results corroborate the conclusion made in the relevant literature that the FDR procedure controls the false discoveries but does not (directly) control the misses (false negatives). Second, the power of the FDR procedure depends primarily on the distribution of the p-values under the alternative hypothesis. The distribution of the p-values under the alternative is determined by the power of the local hypotheses tests that produce the p-values. Third, the p-values must be arbitrarily small so that the FDR procedure can discover all true alternative hypotheses. Forth, the proportion π_0 of true nulls as well as the total number m of hypotheses affect the FDR power: a larger π_0 and/or m decrease FDR power when the local detectors are not powerful enough. Fifth, our results show that when the hypothesis test has low power some of the p-values under the alternative hypothesis are clustered around 0 and the remaining ones seem to be uniformly distributed.

Chapter 7

Evaluation of the large-scale statistical inference framework

We examine the performance of the proposed large-scale statistical inference framework under diverse conditions. Large-scale inference refers to the dimension of the collected data—the number m of p-values—generated by the massive number m (up to hundreds of thousands) of sensors.

7.1 Simulation scenarios

7.1.1 Local conditions

The conditions under which the local detector operates are simulated with parametric distributions: *normal* $\mathcal{N}(\mu, \sigma^2)$, *exponential* $\text{Exp}(\lambda)$, *Rayleigh* $\text{Ray}(s)$, *Rice* $\text{Ric}(s,\ k)$ and central and noncentral *chi-squared* $\chi^2(\vartheta)$. According to [40], the parameters of the distributions under the null and alternative hypotheses (denoted by Θ in **Algorithm 4**) are chosen in [40], so that these distributions resemble each other as closely as possible. Consequently, we use the same local conditions and summarise them in Table 7.1. Their empirical distribution functions are visualised in Appendix C.

The local detector is tuned with the size d of the training data I, the number of bootstrapped samples B, ambient and observation data samples X and Y of length n and l, respectively. In the simulations, we used the same values as those reported in [40]. We list them in Table 7.4 for convenience.

Table 7.1: Local conditions under which the nonparametric detector operates. The parameteric distribution F modells the nominal conditions $\mathcal{H}_0$ and the parametric distribution G models the conditions under the alternative hypothesis $\mathcal{H}_1$. The Kolmogorov-Smirnov test statistic and the p-value are calculated using Matlab for two randomly sampled data sets of length $n = 500$ and $l = 100$ from distribution F and distribution G, respectively.

set-up	F	G	KS test stat	p-value
a	Noncentral $\chi^2(2)$	Central $\chi^2(2)$	0.21	0.0010
b	Ric(1,1)	Ray(1)	0.18	0.0077
c	Ray(1)	Ric(1,1)	0.21	0.0009
d	Ray(1)	$\mathcal{N}(1,1)$	0.28	0.0000
e	$\mathcal{N}(1,1)$	Ray(1)	0.22	0.0005
f	$\mathcal{N}(1,1)$	Exp(1)	0.19	0.0053
g	Exp(1)	$\mathcal{N}(1,1)$	0.18	0.0060

Table 7.2: Parameters of the local nonparametric detector based on bootstrapping and Anderson-Darling (AD) test [40].

training data I	bootstrapped samples	X sample	Y sample
d	B	n	l
10^4	10^3	500	100

7.1.2 Network parameters

The network related input parameters are chosen to represent the most demanding state of the network: a large number of sensors m and large π_0. In fact, π_0 reflects the state of the monitored area but it also depends on the location of the sensors, which is the reason to discuss it here. The sensitivity study conducted in Chapter 6 shows that a large proportion π_0 of true null hypotheses, decreases the FDR power. The achieved power is the smallest when $\pi_0 = 0.9$ and the largest for $\pi_0 = 0.1$ under otherwise equal conditions. We evaluate the large-scale statistical inference performance for $\pi_0 = 0.9$, which is representative for a worst-case scenario in terms of m_0 and m_1, for the reasons as follows. The state of the monitored area can change over time. As a result, π_0 is determined by the natural phenomena that can occur in the field; that is, π_0 is a random variable, not a parameter that can be controlled by the designer. The interest then is on the performance under

the most demanding conditions as for all other values of π_0, the performance is guaranteed to be better. In fact, there will be some degradation of the performance for $0.9 < \pi_0 \leq 1$, but we exclude the two extremes 0 and 1 and consider $\pi_0 = 0.9$ as it is sufficiently representative for worst-case conditions.

7.1.3 FDR parameter

Furthermore, a more restrictive FDR threshold leads to a lower FDR power compared to a larger q^*. Therefore, evaluating the performance at $q^* = 0.05$ is a worse-case scenario compared to $q^* = 0.2$.

7.1.4 Phenomena

We evaluated the performance of the large-scale distributed statistical inference first under a simple set-up in Section 7.2, similar to that in Chapter 6, in which the nominal conditions in the entire monitored area are the same; that is, each of the m sensors operates under the same nominal conditions as the other nodes. The phenomenon under the alternative hypothesis for each sensor is modelled with the same underlying distribution. Such conditions might be unrealistic especially if the area of interest is very large with thousands of sensors. Nevertheless, they allow for understanding and analysing the performance before studying it under more realistic and thus more complex scenarios in Section 7.3.

7.2 Performance under common underlying distributions

We first ilustrate the impact of different variables and tuning parameters on the performance of the large-scale statistical inference and then simulate all the cases listed in Table 7.1.

7.2.1 Total number of sensors m

In the related literature (see [7] for instance) as well as in 6.2.1, some drop in the power of the FDR control procedure is recorded when the total number m of simultaneously tested hypotheses is increased. This effect was studied for all cases listed in Table 7.1 and is exemplified for one of them in Table 7.3. No relevant decrease in the FDR detection power is observed for the examined conditions.

Table 7.3: 95 % bootstrapped confidence intervals for the median $\widetilde{P_d}$ of the FDR power P_d when $\pi_0 = 0.9$ and $q^* = 0.05$. The number of sensors does not have a large impact on power for the studied conditions.

m	F	G	$\widetilde{P_d}$
10K	$\mathcal{N}(1,1)$	Ray(1)	[0.9260, 0.9290]
50K			[0.9255, 0.9282]
100K			[0.9226, 0.9256]

7.2.2 Proportion of true null hypotheses π_0

The performance of the FDR control porcedure at the fusion centre depends on the proportion π_0 of true null hypotheses, see 6.2.2. Therefore, it is of interest to quantify its effect on the achieved power at the FC. Note that $\pi_0 = m_0/m$ is a random variable and as such is not a designer's choice. Furthermore, it is not observable in practice, although it can be estimated.

We have simulated a network of $m = 500$ sensors (since we are interested in the general trend and the impact of m on FDR power is not relevant according to Table 7.3, while the simulation time is sharply decreased for such small m) with a common underlying distribution under the null modelled with a Ric(1,1) distribution. The non-nominal conditions are simulated with Ray(1) distribution, see **b** set-up in Table 7.1. Hence, the observation sample Y obeys Ric(1,1) distirbution for m_0 sensors and for the remaining $(m - m_0)$ sensors Y is sampled from a Ray(1) distribution.

Fig. 7.1 shows that the detection power of the FDR control procedure decreases as π_0 increases. The same observation was made for all scenarios from Table 7.1. The magnitude of the observed effect depends on the particular underlying ($\mathcal{H}_0$ and $\mathcal{H}_1$) distributions and is the largest for the $\mathcal{N}(1,1)$ vs Exp(1) (F vs G, case **f** in Table 7.4) from those examined. In fact, for all but this latter scenario, the decrease in power is about 10% when π_0 is increased from 0.1 to 0.9. The decrease in power for the $\mathcal{N}(1,1)$ vs. Exp(1) case is by more than 60%, which result we analyse in Section 7.2.4.

7.2.3 FDR threshold q^*

The magnitude of the impact of the FDR bound on the achieved FDR power is examined for predominating number of null hypotheses $\pi_0 = 0.9$ when the set-up is **f** and **g**, Table 7.1: $\mathcal{N}(1,1)$ vs Exp(1) and Exp(1) vs $\mathcal{N}(1,1)$. The FDR power for the former is low—median $\widetilde{P_d} \approx 33\%$—and for the latter,

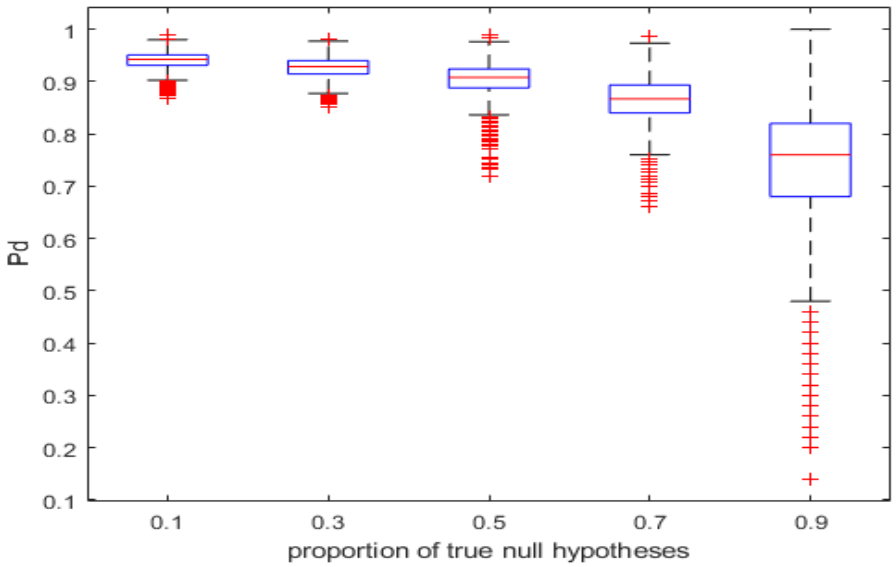

Figure 7.1: Detection power for different proportions of true null hypotheses π_0. The conditions under the null hypothesis $\mathcal{H}_0$ are modelled with Ric(1,1) distribution at each sensor. The conditions under the alternative hypothesis $\mathcal{H}_1$ are modelled with Ray(1). FDR power is the largest when the number m_1 of new phenomena that occur in the sensor network is the largest.

it is high—$\widetilde{P_d} \approx 90\%$. A higher threshold q^* allows for a larger number of discoveries to be made: in the case of $\mathcal{N}(1,1)$ vs Exp(1) an increase of more than 50% is observed, whereas in the Exp(1) vs $\mathcal{N}(1,1)$ the improvement is less than 10%.

It was shown in 6.2.3 that a less stringent cut-off threshold q^* yields higher FDR power, but a larger number V of false discoveries. In both cases plotted in Fig. 7.2 and Fig. 7.3 the larger detection probability is achieved at the cost of a higher number of false discoveries, namely 1 vs 10 on average. Furthermore, the results show that V depends on the underlying distributions as well (compare results from $\mathcal{N}(1,1)$ vs Exp(1) to results from Exp(1) vs $\mathcal{N}(1,1)$ cases).

The total number of sensors (hypotheses) m impacts the number of false discoveries too as expected. When $m = 50\text{K}$ and $q^* = 0.05$, the median of V is much higher: 275 for the **f** scenario (corresponding to Fig. 7.2) and 240 for the **g** conditions (corresponding to Fig. 7.3) in Table 7.3.

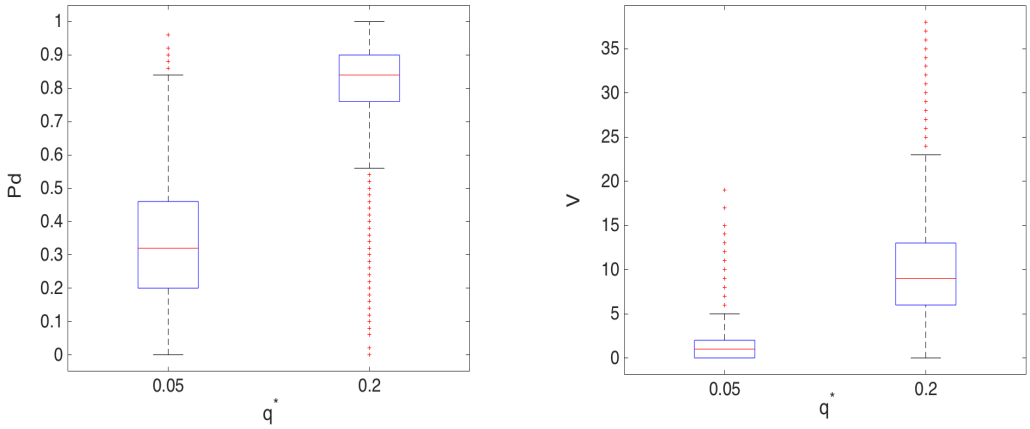

Figure 7.2: Effect of the predefined FDR threshold q^* on the FDR power P_d. $\mathcal{N}(1,1)$ vs Exp(1) is simulated. Larger threshold leads to a higher detection power but also to an increased number of false discoveries on average.

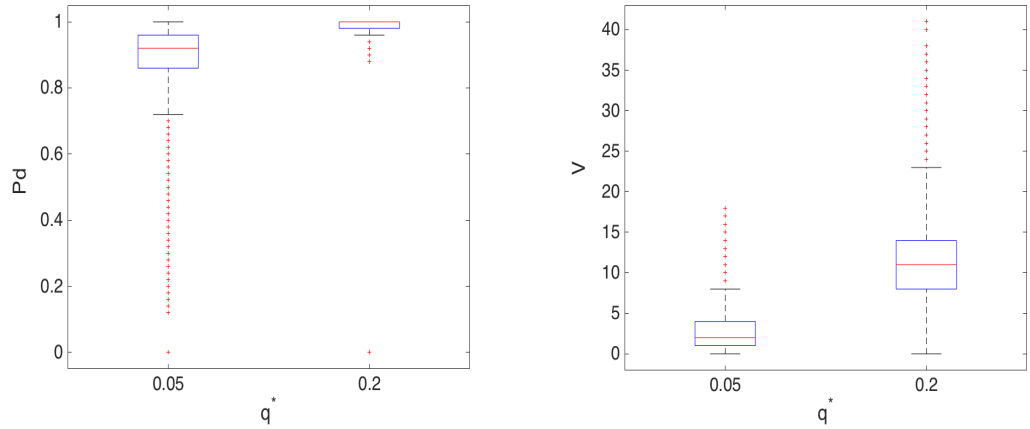

Figure 7.3: Effect of the predefined FDR threshold q^* on the FDR detection probability P_d. A scenario of Exp(1) vs $\mathcal{N}(1,1)$ is simulated. Larger threshold leads to a higher detection power at the cost of an increased average number V of false discoveries.

7.2.4 Evaluation results and analysis

Figure 7.4 shows the performance of the proposed detection framework for a large-scale IoT network comprising $m = 50,000$ sensors. The proportion of true null hypotheses is set to $\pi_0 = 0.9$, which means that only 10% of all sensors observe a new phenomenon (a departure from the nominal conditions); the remaining 90% do not observe any change. The FDR bound is set to $q^* = 0.05$. This is the most demanding set-up. Each box plot represents a different scenario (see Table 7.1): nominal conditions (under $\mathcal{H}_0$) and conditions under the alternative hypothesis $\mathcal{H}_1$. Despite that this is the worst-case scenario among all studied, the median of the achieved power is above 75% for all but one case. In the context of this result it is worth noting that when only 10% of the null hypotheses are true (changing a single input simulation parameter, $\pi_0 = 0.1$), the mean and median of the FDR power is larger than 90% for all simulation scenarios from Table 7.1.

To understand the former of the obtained results (power lower than 75% for $\mathcal{N}(1,1)$ versus Exp(1) case) recall that two important conclusions are made in Section 6.3 regarding the power of the FDR procedure. It depends on the proportion π_0 of true nulls; the magnitude of the effect is determined by the power of the local detectors. In particular, when the power of the local detectors is 1 (that is, the p-values are arbitrarily small), π_0 does not have any impact on FDR power. However, when the detection power of the sensors decreases, the proportion π_0 has an immediate, negative effect on the attained power.

Therefore, to explain the "outlier" in Fig. 7.4, namely $\mathcal{N}(1,1)$ versus Exp(1) scenario, we examined the p-values of this and the other local conditions from Table 7.1 as well as the detection power of the sensors. The acceptance ratio of the nonparametric detector was calculated for $\alpha = 0.05$ significance level. For each scenario the acceptance ratio was averaged over 10^4 Monte Carlo realisations. This step was repeated 10^3 times (in total 10^7 MC simulation runs), so that we obtained $1,000$ values for the mean of the acceptance ratio of the non-parametric local detector. The 95% bootstrapped confidence intervals were calculated based on these $1,000$ points. These are summarised in Table B.1. The distribution of the p-values from the 10^7 realisations is shown on Fig. 7.5 and Fig. 7.6. Clearly, the p-values under $\mathcal{N}(1,1)$ versus Exp(1) case are larger than the other three cases plotted there (for which three cases most of the p-values are centred at 0). Recall that the FDR power is high as long as the p-values of the true alternative hypotheses tend to 0, Section 6.3.

It is also important to note that when all hypotheses are true alternatives ($\pi_0 = 0$), FDR power is determined by the power of the non-parametric

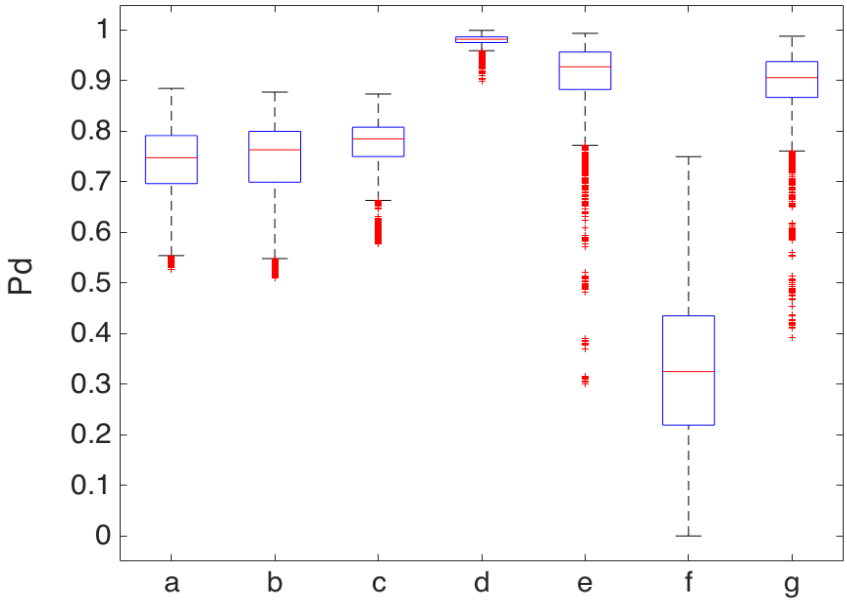

Figure 7.4: FDR power for a scenario, where the nominal conditions are modelled with the same distribution for all sensors. Similarly, the observed phenomenon is modelled with a single distribution for all nodes in the network. The cases listed in Table 7.1 are simulated. The total number of sensors is $m = 50$K, from which only 10% observe a departure from the nominal conditions ($\pi_0 = 90\%$). The FDR threshold is set to $q^* = 0.05$. Despite that this is *a worst-case* scenario, for all studied cases except for (**f**), $\mathcal{N}(1,1)$ vs Exp(1), the median of the detection probability is above a 75% level.

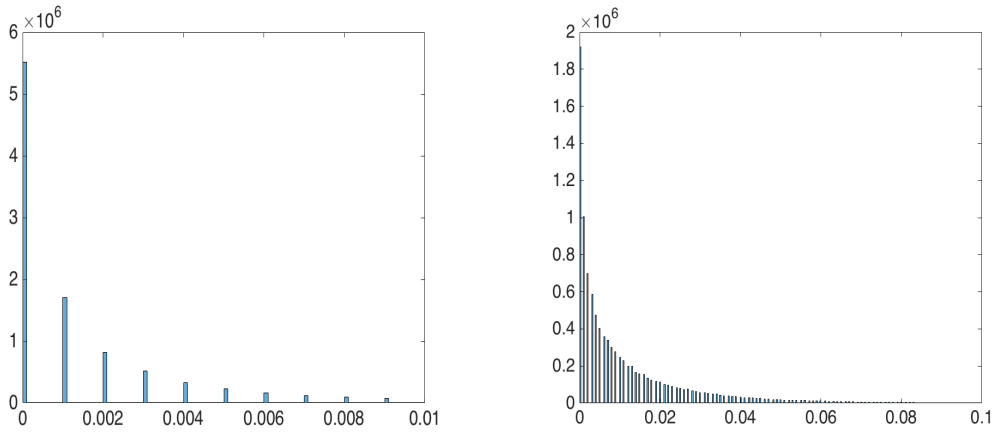

Figure 7.5: Histogram of p-values from 10^7 Monte Carlo realisation of the local detector when: (*left*) the null hypothesis is that samples X and Y obey the same Exp(1) distribution whereas observation sample Y actually obeys $\mathcal{N}(1,1)$, row 1 in Table B.1; (*right*) $\mathcal{H}_0$: X and Y obey $\mathcal{N}(1,1)$, whereas Y is sampled from Exp(1), row 2 in Table B.1.

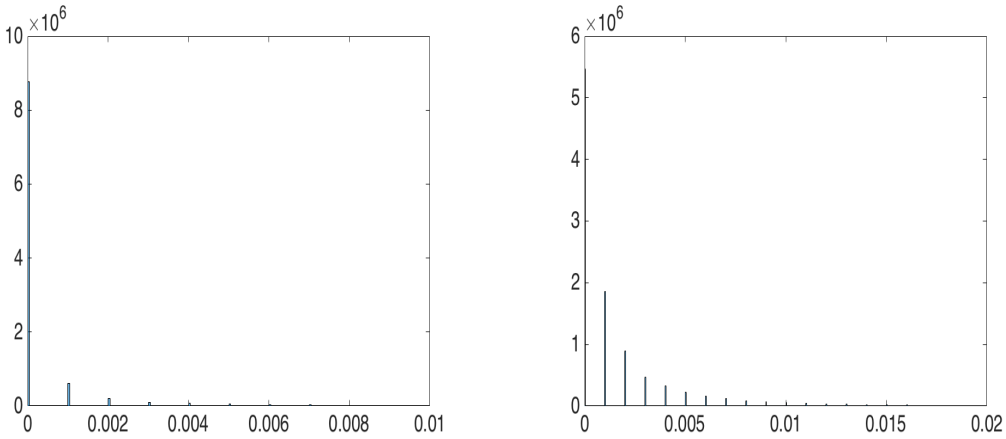

Figure 7.6: Histogram of p-values from 10^7 Monte Carlo realisation of the local detector when: (*left*) the null hypothesis is that samples X and Y obey the same Ray(1) distribution whereas observation sample Y actually obeys $\mathcal{N}(1,1)$, row 5 in Table B.1; (*right*) $\mathcal{H}_0$: X and Y obey $\mathcal{N}(1,1)$, whereas Y is sampled from Ray(1), row 6 in Table B.1.

detectors. In effect, for the examined $\mathcal{N}(1,1)$ versus Exp(1) and $\pi_0 = 0$, $P_d \approx 95\%$. Nevertheless, when π_0 increases, FDR power decreases. The decrease is in function of the power of the local detectors and the number of true null hypotheses m_0 (or π_0). Under a large π_0, $\mathcal{N}(1,1)$ versus Exp(1) exhibits the lowest power among the other scenarios in Fig. 7.4 since the power of the nonparametric detector is the lowest for the $\mathcal{N}(1,1)$ versus Exp(1) case, see Fig. 7.5 and Fig. 7.6.

7.3 Performance under multiple different underlying distributions

The versatility of the proposed large-scale statistical framework stems from the fact that its practical application is not confined to scenarios where the distribution under the null is common to all sensors. In fact, it can be different at each sensor. Recall that we do not make any assumptions about the observed nominal conditions nor about the distribution of the phenomenon under the alternative hypothesis. Furthermore, while the nominal conditions (i.e., the probability distribution under the null $\mathcal{H}_0$) must remain the same throughout the lifetime of the sensors, the probability distribution under the alternative hypothesis $\mathcal{H}_1$ can change over time (i.e., between observation periods) at each sensor. Therefore, as a next step, we study a more complex set-up in which the nominal conditions as well as phenomena observed at each group of sensors are different from the remaining groups. Here, group can consist of a single sensor or of all the m sensors.

To exemplify the performance of our large-scale statistical inference concept under more realistic IoT conditions we simulated a network comprising $m = 50$K sensors, which we divided into 5 groups. Each cluster consists of 10K sensors. The conditions in these groups were modelled with the cases listed in Table 7.1 except for Ray vs Ric and $\mathcal{N}(1,1)$ vs Ray(1), (**c** and **e** scenarios), which were not included. In each group only 10% of the sensors are observing departure from the nominal conditions. For each group of 10K sensors (consider Noncentral vs Central χ^2, case **a**, for instance) the nominal as well as alternative conditions are modelled with the same parametric distribution (Noncentral χ^2) for 9K of the sensors in the group. The nominal conditions for the remaining 1K sensors are also modelled with the same distribution (Noncentral χ^2) but the conditions under the alternative with a different probability model (Central χ^2 in **a** case).

The histogram of the probability distribution of the FDR power over the 10^4 MC realizations is shown in Fig. 7.7.

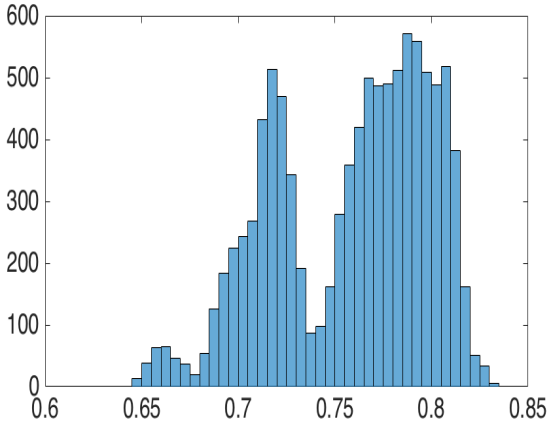

Figure 7.7: FDR power when there are 5 different clusters of sensors. The underlying distributions within a cluster are the same for all sensors but differ between clusters. The performance is studied under worst-case conditions: $\pi_0 = 0.9$ in each cluster. The FDR threshold is set to $q^* = 0.05$.

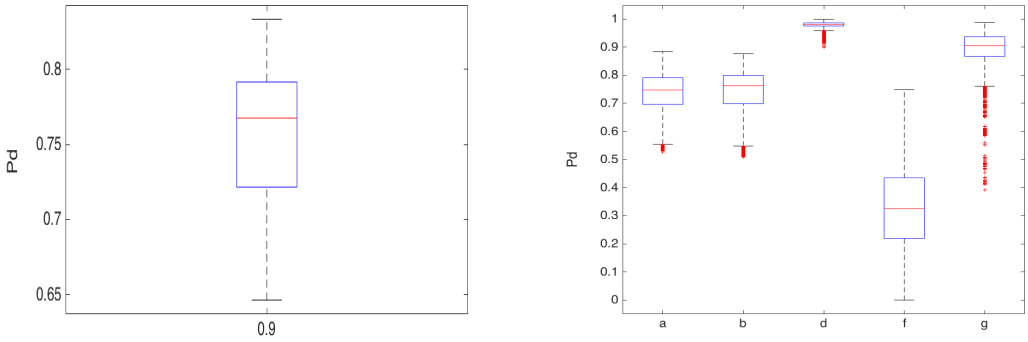

Figure 7.8: Boxplot of FDR power for the conditions studied in Fig. 7.7 (*right*). The power for each cluster (*left*) .

In practice, the detection probability of each cluster depends on the power of the local detector for the specific probability model as well as the proportion of true null hypotheses within the cluster, and thus can be different for each group. We used a common $\pi_0 = 0.9$ proportion of true nulls in each group because of comparison reasons and ease of interpretation. The histogram plotted in Fig. 7.7 can be explained by looking at Fig. 7.8. The scenario for which the simulation results are presented in Fig. 7.7 is a combination of 5 of the scenarios from Table 7.1 and the peaks in the histogram roughly correspond to the composition of the medians plotted in Fig. 7.8.

We see the coupled effect of the different groups. The box plot of the FDR power from all realisations in Fig. 7.8 *right* is compared to the box plots in Fig. 7.8 *left*. Clearly, the median of the power from the 10^4 MC realisation reflects the combination of the medians obtained at each cluster.

The distribution of the p-values of the alternative hypotheses from a single simulation run are plotted in Fig. 7.9 for ilustration. The power recorded from this MC realisation is 0.71.

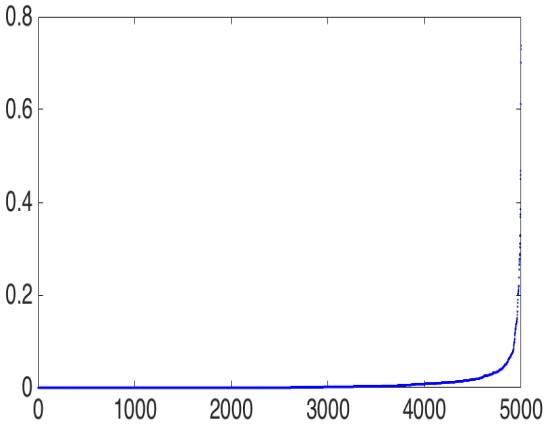

Figure 7.9: The first $p_i \approx 0, i = 1, ...3500$ rank ordered p-values from all alternative hypotheses are correctly discovered by the FC, but the remaining p-values are large and the FDR procedure fails to reject them. The number of false discoveries is $V = 159$ for this simulation run.

7.4 Conclusions and future research prospects

Depending on the particular application, the requirement on the attained FDR power can be very high. The results and analysis in the previous sections demonstrate that the power of the local detector determines to a large extent the power of the suggested statistical inference framework. We discuss two approaches for improving the local detector power and consequently the performance of the studied statistical inference concept.

7.4.1 Sample size

One suggestion made in Section 6.2 is that the size of the data sample can improve the detection power of the non-parametric local detector and consequently FDR power too. We empirically examined the statistical infer-

ence performance under different values of the local detector parameters, Table 7.4. We looked into different B, l and n. For $n = 2500$ and $m = 500$, FDR power is 100% in all $MC = 10^4$ realizations under all local conditions in Table 7.1 when $\pi_0 = 0.9$, $q = 0.05$ and $m = 50$K. The performance was also studied for an even larger number of sensors—$m = 100K$—under Noncentral $\chi^2(2)$ versus Central $\chi^2(2)$, and Exp(1) versus $\mathcal{N}(1,1)$ probability models. Despite this massive number of sensors, FDR power is 100% due to the increased power (100%) of the local detectors. This improvement is a direct result of the increased sample sizes.

Table 7.4: Parameters of the local nonparametric detector based on bootstrapping and Anderson-Darling (AD) test [40].

training data I	bootstrapped samples	X sample	Y sample
d	B	n	l
10^4	10^3	2500	500

Since the proposed statistical concept can be used to delineate areas where a new phenomenon in a sensor network is observed (consider for instance the problem of detecting and localising flooding occurring in an isolated area of an agricultural field), it is relevant and interesting to look into the proportion of false discoveries as they could impact the correct localisation of the event. We make an implicit assumption that the location of the sensors in the network is known or it could be accurately estimated.

Table 7.5: The median $\widetilde{V}$, mean $\bar{V}$, standard deviation of V, minimum and maximum value of V for Ray(1) vs Ric(1,1) and $m = 50$K sensors.

sample size	$\widetilde{V}$	$\bar{V}$	std deviation V	max(V)	min(V)
n=2500, l=500	262	274.3134	111	861	0
n=500, l=100	189	209.3147	103	579	0

Table 7.5 provides specific statistics regarding V for a Ray(1) vs Ric(1,1) conditions, $m = 50$K sensors and $\pi_0 = 0.9$ ($m_1 = 5,000$ sensors observe departure from nominal conditions). The sample length of the ambient X and observation Y data sets at the local detectors is set to: $n = 500(2500)$ and $l = 100(500)$, respectively. Note that minimum and maximum value of V are outliers and the main bulk of data points from the $MC = 10^4$ realizations

is roughly in the $(150, 350)$ interval (see Fig. 7.10), which compared to the total number m_1 of sensors with true alternative can be regarded as relatively small. Furthermore, as long as the false discoveries are not clustered together, their impact might not be large or mechanisms that explore the nature of the particular IoT application can be employed to distinguish them from the true discoveries.

Another relevant observation is that the sample size determines the power of the local detector but also has an impact on the number of false discoveries V. Larger sample sizes increase the power of the statistical inference framework but do not lead to smaller number of false discoveries, Table 7.5; that is, a larger ambient and observation data samples might not make easier the estimation of the boundaries of the area where a phenomenon occurs.

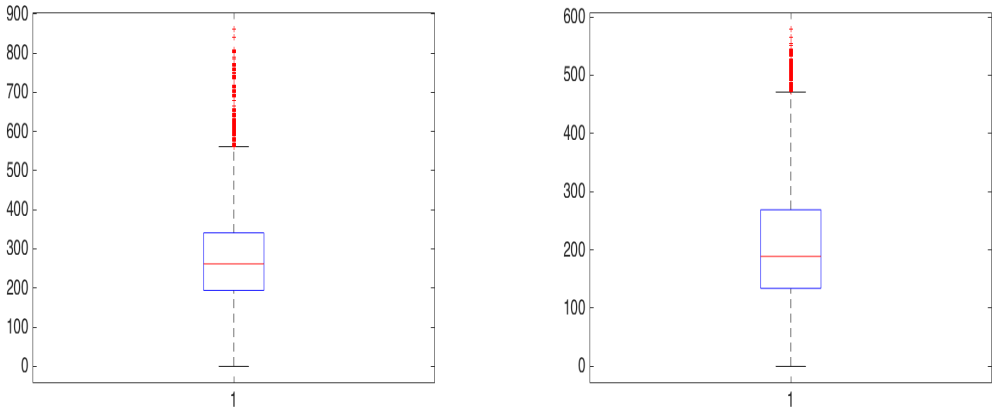

Figure 7.10: Boxplot of the number of false discoveries V for large (*left*) and small (*right*) data samples.

Note that under the local detector design, Algorithm 2, these larger sample sizes might translate into longer sampling time during observation and consequently higher energy consumption. Commonly, sensor nodes are subject to very stringent power requirements [41], [42], [43] and sensor network design is normally accomplished under strict energy limitations (see for instance [44]) despite that recent studies also look into cases when energy harvesting [45] can be integrated into the network. Therefore, in certain applications where sensors operate with batteries and the network is expected to remain functional for several years and be energy-efficient, such an approach might not be feasible.

7.4.2 Design

There are few immediate observations we make about the design of the local detector, Algorithm 2: (i) large training data set I, which is not fully explored; (ii) implementation of bootstrap principles despite large I, while the main reason for bootstrapping usually is lack of data; and (iii) resampling is done at each decision point despite that nominal conditions remain the same.

According to Algorithm 2, the objective of the non-parametric local detector during training is to collect a large data sample I (potentially during an extended period of time). The length of the training data in the simulations is $I = 10^4$. This data sample I is used at two stages: during training to obtain the EDF of the test statistic and during hypothesis testing to emulate sampling from nominal conditions. Note that in none of these two phases the entire data set is used. Storing a large data set might have implications for the memory requirements too.

An important implication of the non-parametric local detector design, Algorithm 2, is that resampling from I is performed at each decision point. In the scientific literature, sensor networks are assumed to consists of a massive number of cheap devices with limited energy that must operate during prolonged periods of time (in terms of years) [41], [42]. Therefore, a fundamental principle in sensor networks design is energy conservation [43] as noted earlier. Note, however, that the resampling performed at the local detector means increased computational burden and a constant draining of battery due to periodic (at each decision point) repetition of this step. Furthermore, the nominal conditions are sampled at each decision point despite that they are assumed to remain the same.

Another relevant observation that became apparent when analysing the results in Fig. 7.4, Section 7.2.4 is the variability of the non-parametric local detector power. This variability of the results naturally transfers into variability of the results regarding the inference framework. A closer look into Algorithm 2 suggests two sources of randomness: the sampling of nominal conditions and sampling during observation. To elaborate further on the random aspect, consider the two extremes when sampling from nominal conditions: all data points come from the tails of the distribution and all sampled data points are uniformly distributed along the complete EDF.

One potential way we saw for overcoming the aforementioned side effects of resampling is the use of a single EDF with the main expected outcomes:

- decreased variability of results,
- decreased (periodic) computational burden,

- avoidance of energy depletion due to repetitive sampling from nominal conditions.

Learning a single EDF can eliminate the first source of randomness and, as long as the learnt EDF is sufficiently representative for the observed probabilistic model, it can yield better detection performance results, we conjecture. Such design will avoid the repetitive sampling under nominal conditions. This can in fact solve the evident energy problem of Algorithm 2 and lead to more optimal battery use and better energy balance.

Chapter 8

Relevant prior art

We centre on and present a chronological overview of the contributions that employ the principles of the FDR control procedure in wireless sensor networks. Common to the work of Ermis and Saligrama and Ray and Varshney is that the target to be detected is modelled to emit a signal that decays with distance. The primary difference between the focus of the two research lines is in the scenarios considered. Ray and Varshney study the traditional set-up with a single target within the range of all nodes. Ermis and Saligrama, complement the conventional focus with detecting multiple events, which can occur in different parts of the network and which are in the vicinity of only a small subset of sensors.

8.1 Distributed FDR for multitarget detection

Ermis and Saligrama address a *multitarget* distributed detection problem – a sensor network where multiple phenomena can simultaneously occur [20]–[24], [26]. The main assumptions the authors make are that only a small number of sensors are in the vicinity of a phenomenon and the sensors measure independent—conditioned on the hypothesis—information.

The authors study the traditional distributed detection scenario with a central node. The configuration is not parallel; instead, a sensor decision is broadcast to all nodes in the network (fusion centre and remaining sensors). There is no feedback channel from the central node to the sensors. The primary networking constraint is the cost induced by communicating test decisions from the sensors to the network. The communication cost accounts for limited bandwidth and energy. Each sensor can use 1 bit for communicating its local binary decision.

Algorithm 6 Distributed FDR procedure by Ermis and Saligrama

At each sensor:

Step 1. Calculate the p-value of the test statistic:

$$p_i(X) := \int_X^\infty f_0(t)\,dt = 1 - F_0(X), \quad i = 1, \ldots, m.$$

Step 2. Compare the calculated p-value p_i to the current common global threshold γ_j, which is given by:

$$\gamma_j = (s_j/m)\, q^*,$$

where q^* is the FDR constraint, s_j is the updating constant, $s_j \in \mathbb{N}$, and j indicates the communication round. Initially, before the first round, $j = 0$ and $s_j = s_0 = 1$. Hence, the initial global threshold is set to

$$\gamma_0 = (1/m)q^*.$$

Step 3. Declare the observation as significant if

$$p_i \leq \gamma_j$$

and broadcast this decision to the entire network.
Assume there are l_{j+1} broadcast decisions.
Step 4. Update the global threshold to

$$\gamma_{j+1} = \frac{s_{j+1}}{m} q^*$$

at each sensor, which has *not* yet broadcast its decision, where $s_{j+1} = s_j + l_{j+1}$.
Step 5. Repeat **Steps 3** and **4** until there are no more sensors that declare their observations as significant.

At the fusion centre:

Step 6. Let the total number of sensors that have declared their observation as significant be m^* after the last communication round.
Decide the alternative hypothesis for those $\mathbf{m}^*$ sensors.

The problem is mathematically defined as follows:

$$\begin{aligned} \min \quad & E\left(T\right) \\ \text{s. t.} \quad & E\left(V \,/\, R\right) \leq q^{*}, \\ & \sum_{i,t} c\left(u_i(0), \ldots u_i(t)\right) \leq C, \end{aligned}$$

where T is the number of misses (or type II errors), V is the number of erroneous discoveries, R is the total number of discoveries, c is the cost of communicating a decision and C is the total communication constraint.

Ermis and Saligrama address the problem by developing a distributed FDR control procedure in [20]–[22], which we summarise in **Algorithm** 6. The proposed method is a sequential algorithm that linearly increases the FDR threshold. The algorithm can potentially save some energy (at the cost of induced delay) since the nodes broadcast their decision only if an observation is declared significant. Therefore, Algorithm 6 can be viewed as a special class of distributed detection with censored observations [27].

Ermis and Saligrama note in [20] that the performance of the FDR procedure "does not depend on the probability distribution under the alternative hypothesis" [20] and that there is no control on the miss rate (false negatives or type II errors). Based on this the authors conclude that the detection power at the fusion centre can be poor and point out that the "FDR procedure performs best when the p-values of the data that comes from $\mathcal{H}_1$ are clustered near 0" [20], [21]. To overcome this problem, a transformation of the p domain is introduced [20]–[22] under the hypothesis that the distributions of the observations under $\mathcal{H}_0$ and $\mathcal{H}_1$ are known. The transformation is applied on each p-value prior to running the FDR algorithm. The authors assure that the the p-value transformation increases the probability of declaring an observation as significant, see Fig. 4 in [21].

In [20] and [21] a boundary detection problem is tackled. Therefore, in addition to the Bonferroni procedure, the distributed FDR control algorithm is compared to another algorithm designed for boundary problems. The comparison is in terms of detection performance and communication cost.

Unlike earlier work [20]–[22], in [24] Ermis and Saligrama model the observetions of the sensors without a target in their range as the sum of the decayed signals from far away targets plus noise. In the ideal case, the sensors would observe only noise; that is, the decayed signals are considered undesired disturbance in the formulation of [24]. The undesired distortion of the known noise model is unknown. Hence, the observed model is unknown and different at each sensor. Thereby, an exact model cannot be constructed

under the null hypothesis [24]. It is shown that the p-transformation is robust to perturbations in the observed model. The established robustness refers to the increase in the miss rate when the perturbations of the ideal signal model increase. On the other hand, robustness is interpreted as the ability of a family of distributions that have the same variance to remain close to each other after the application of the aforementioned transformation. Results and correspondingly conclusions similar to [20]–[22] are reported in [24].

In [25], a dynamic version of the distributed FDR method [20] is proposed. The false discovery rate in the dynamic and distributed FDR is constrained by q^*, rather than by $(m_0/m)\, q^*$: FDR $\leq q^*$ instead of FDR $\leq (m_0/m)\, q^*$. Originally, the idea of adaptive FDR control was proposed and explored by Benjamini and Hochberg in [17], where it is demonstrated that the dynamic adjustment of the pre-chosen q^* can improve the power of the procedure.

To clarify this idea we draw the reader's attention to one of the inherent characteristic of the Benjamini and Hochberg's FDR procedure, namely that the false discovery rate "autonomously" tunes itself to the current state of the system. Here, system state is given by the proportion of true null hypotheses $\pi_0 = m_0/m$ or the number of sensors, which do not observe a target in their sensing range. Hence, when the number of true null hypotheses is small compared to the total, the false discovery rate will be small and likewise when the number of true hypotheses m_0 approaches the total m, the FDR will be (at most) q^*. In other words, the larger the m_0, the higher the number of (allowed) erroneous discoveries. In presence of true alternative hypotheses, the FDR is controlled at level $\pi_0\, q^*$. In case this π_0 proportion can be estimated, the algorithm can be modified so that control is always performed at a level q^* independent of the actual system state. This will allow for a larger number of discoveries or equivalently larger power, see [17].

We note that the simulation results reported in [25] depict a lower empirical error for the modified compared to the original procedure. However, it should be clear at this point that by changing the FDR threshold q^* as suggested in [17] and [25], the procedure allows for a potentially higher number of false discoveries to be made and thus not only for a higher detection power but for a higher error rate too. In short, the larger number of discoveries is at the cost of a larger number of type I errors. Therefore, the type I errors with the original procedure are at most as high as of the modified and not vice versa. The authors do not explicitly clarify in [25] what meaning they attribute to the term *empirical error*; perhaps, they refer to the total number of errors (type I an type II) [26].

The possibility of having multidimensional observations at each sensor is investigated in [26]. A transformation of multidimensional to a scalar test statistic is proposed. The transformation is based on the Radon-Nikodym

theorem (volumes of level sets of the likelihood ratio function). The effect of object density on the performance of the FDR procedure that uses the proposed transformation is studied. Further, the impact of the attenuation coefficient on the error rate and communication cost is examined.

In [21] the authors depart from the multitarget detection and look at the problem of detecting and localising a *single* object that emits a signal with unknown power. The signal is globally observable, but it is assumed that the signal power of the emitter decays rapidly with the distance. Each sensor declares that the target is present if it is within its range; otherwise, the target is considered absent in this formulation. The (local) false alarm probability therefore is defined as the probability of a sensor declaring the presence of the object when the object is outside its range. A miss occurs when the sensor declares no target present when the target is within its immediate vicinity. The assumptions made are: only the fusion centre knows the (exact) locations of the sensors (expressed as the distance between the fusion centre and the node) and two bits in total can be used by each sensor. The (approximate) location of the object is determined at the FC by averaging the distances between the sensors with significant observations and the FC. Here, **Algorithm** 6 is applied (without the transformation of the p-values proposed in [20]–[22]) to determine which observations are significant. The authors comment on the accuracy of such location estimation and note that the sensor locations can be weighted according to the exhibited signal-to-noise ratio.

8.2 Distributed FDR for traditional single target settings

Ray and Varshney [28]–[33] apply the distributed FDR detection concept of Ermis and Saligrama under the same constraints set in [20]–[26], namely communication and energy cost but to a single-target scenario. A primary difference between these two groups of studies is that the decision boundary of the distributed FDR algorithm in [20] is found from the left, starting from the threshold q^*/m and updating the global threshold according to **Algorithm** 6. In contrast, in [28] the search is performed from the right, starting from the FDR threshold q^* (denoted by γ in [20]–[33]) and updating it accordingly. In the $j+1$-th round with right update, for instance, the threshold is given by $\gamma_{j+1} = (m - s_{j+1})/m$, where s_{j+1} is calculated as in Algorithm 6. The benefit of the latter is that the p-value with the largest index can be found and therefore the largest probability of detection can be

attained. This is the intended behaviour of FDR (see Algorithm 3) where the largest $p_{(i)}$ satisfying (4.1) is sought. It is argued in [26] that under the p-transformation the direction of the search does not have any effect on the observed power "first-crossing", which refers to updating from the left, and "last-crossing" or updating from the right, "lead to same performance with high probability" [26].

As mentioned earlier, Ray and Varshney study distributed detection of a *single* phenomenon [28], [29], [32], [33]. Additional assumptions made are: a total of m sensors are uniformly distributed in a region of interest (a square in [28], [29] and a disk in [32], [33]). When a target is not present in the detection area, noise n_i is observed at sensor $i, i = 1, \ldots, m$. The noise is modelled as a random variable that obeys a standard normal distribution. The target emits a deterministic signal with amplitude a. The *signal* to be detected is modelled at each sensor through an amplitude $a_i > 0$, which depends on the distance between the target and the i-th sensor. The local hypothesis test at the sensors in this case is:

$$\begin{aligned} \mathcal{H}_0 &: s_i = n_i, & n_i &\sim \mathcal{N}(0,1) \\ \mathcal{H}_1 &: s_i = a_i + n_i, & a_i = \sqrt{P_i}, &\quad i = 1, \ldots, m \end{aligned}$$

where P_i is the received power under the assumed isotropic power attenuation model in [28], [29], [32], [33]. In other words, the signal is modelled as a continuous random variable that follows a normal distribution with mean a_i and variance 1.

In addition, it is assumed that the local decisions are broadcast and the communication medium is error-free. Moreover, a general setting is considered in which the sensors and target locations are unknown to the FC. The authors argue that the Chair-Varshney rule [34] cannot be applied under such assumptions and a new framework for global decision fusion is needed. Indeed, recall that the Chair-Varshney rule is optimal given that the local probability of false alarm and probability of detection at each sensor are known and conditional independence between the observations acquired by the sensors can be assumed. In the aforementioned setup however, the local probability of detection is unknown because of unknown location of the nodes and target.

A difference between the scenarios considered in [28], [29] and [32], [33] is that in the former work, the focus is on the classical distributed detection setting, where all sensors are within the finite radius of influence of the target. In the latter work, the authors recognise the relevance of other realistic cases where the sensors might not be receiving the same target signal due to different radio propagation conditions and/or to the sensors limited sensing

range. In such scenarios, the deployed nodes are divided into two categories in the model: a small fraction of sensors that are in the finite radius of influence of the target and that receive an identical signal and the remaining nodes, which sense only noise. According to [32] such a model reflects detection in chemical and oil fields or detection of electromagnetic and acoustic signals.

The global decision method at the FC proposed in [28], [29], [32], [33] is a counting rule based on the total number of discoveries Δ and a system threshold T. The null hypothesis $\mathcal{H}_0$ at node i is declared true or false by the distributed FDR procedure based on a corresponding p-value and a dynamic cutoff, see **Algorithm** 6. The global decision—target present ot target absent—is determined by the total number of discoveries and the system threshold T as follows:

$$\Delta = \sum_{i=1}^{N} I_i \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} T,$$

where I_i is the local binary decision collected from sensor i (a hard decision about the validity of the null hypothesis).

The authors derive mathematical results concerning the false alarm and detection probabilities under the aforementioned assumptions, which are verified by simulation results. The FDR control parameter q^* and the global threshold T (which corresponds to a total number of discoveries) must be chosen in order to attain a certain system-level false alarm while maximasing the system-level detection probability. Due to analytic and computationally expensive calculations, the authors [32] find the optimal values of these parameters by maximasing the deflection coefficient[1], which is defined by [19]:

$$d^2 = \frac{(\mathbb{E}[T; \mathcal{H}_1] - \mathbb{E}[T; \mathcal{H}_0])^2}{var(T; \mathcal{H}_0)}.$$

However, the maximisation of the deflection coefficient can guarantee optimal global performance if the statistic (in this case the total number of discoveries Δ) is Gaussian. It is shown in [33] via simulations that for non-asymptotic conditions, the maximisation of the deflection coefficient does not yield the expected optimal q^* (γ in [28]–[33]) value and correspondingly maximum detection probability. The authors of [33] propose to use instead the Kolmogorov-Smirnov distance test for finding the FDR threshold q^*. The two cdfs in the Kolmogorov-Smirnov test in [33] correspond to the distribution

[1]Recall that the larger the deflection coefficient, the better the performance of the detector is.

function of the count statistic Δ under the global (at the FC) hypotheses G_0 (target absent) and G_1 (target present). The FDR threshold q^* is determined by maximasing the difference between the two cumulative functions.

In addition to the problem of fusing the decision statistics from the sensors at the central node, the authors [33] consider the problem of Byzantine attacks. It deals with the presence of malicious sensors that aim at interfering with the decision process at the FC by sending modified decision statistics.

A comparison between the fusion decision method in [33], [29], [32] and the Neyman-Pearson formulation with identical decision thresholds (p-value decision threshold for a given local probability of false alarm) at each sensor is performed. The results demonstrate the superiority of using the FDR control procedure.

The FDR approach proposed in [28], [29], [32], [33] is applied to radar detection in clutter (undesired background reflectors [35]) and noise using a *multitarget* formulation.

The differences between the traditional approach to radar signal processing and the method in [30], [31] are summarised next. The former tests each range cell separately at a predefined false alarm probability [31]. When the background statistics are known, the detector design follows the Neyman-Pearson formulation. Otherwise, the neighbouring range cells are used for estimating the clutter and noise. In contrast to this approach, the FDR principle is used in [30], [31] to simultaneously estimate the false discovery rate in m range cells arranged in an annular region. The problem is of detecting Swerling I target in the noise and clutter obeying a Gaussian distribution. A more complicated clutter model is discussed in [31] too, but the authors recognise the difficulties in applying their mathematical methodology to such a case and as in the simple clutter case relay on asymptotic results and simulations to evaluate the system-level false alarm and detection probabilities.

Furthermore, the ideas of the adaptive FDR control described in [17] are incorporated in the radar detection in clutter and noise framework in [31]. Moreover, a hybrid detector is devised too. It combines the FDR and Neyman-Pearson (NP) based approach to improve detection performance in low SNR conditions and sparse targets. Specifically, a logical "OR" operation is performed between the decisions of the FDR algorithm and the NP-test result at each cell in order to determine the presence of a target there.

8.3 Discussion

The main advantage of the distributed FDR procedure proposed in [20]–[26] and later on adopted in [28], [29], [32], [33] is that it avoids communicating the p-values and allows sending a hard decision through a single bit. This benefit comes at the cost of increased latency, which is a random variable that depends on the total number of sensors m and the current p-values. The latency is at least 1 and can be up to m communication rounds. The duration of the communication round is implementation specific.

A potential problem of the distributed FDR control [20] is a sensor failure [41], [42]. It is assumed that the total number of nodes m is known at each sensor. An implicit assumption made is that m remains the same along the lifetime of the network. This assumption has an immediate impact on the algorithm as m is used at each decision cycle of **Algorithm** 6 (Step 2 and Step 4). It can affect the overall performance as long as sensor failures cannot be instantly detected and their proportion is not small.

The communication protocol to be used for feeding the network with sensor decisions is not explicitly considered by Ermis and Saligrama nor Ray and Varshney. Implicitly, sensor decisions are broadcast. The challenges of the wireless communication channel [47] are not addressed either and error free transmissions are considered instead.

In addition to devising a distributed version of the FDR control procedure, Ermis and Saligrama exploit its inherent potential. In effect, FDR control lends itself to multiple phenomena detection and localisation. First, recall that the FDR procedure has been extensively applied to genomics studies. In this context, the FDR algorithm allows identifying genes with different expression levels in healthy individuals and patients or symptom measures that are different under different treatments. The identification in such studies naturally translates into localisation in wireless sensor networks, where the sensors with true alternative hypthesis can be used to determine the area in the network where (different) phenomena occur. Note that for such purposes the location of the sensors must be known, which in contrast to multiarray studies, where each gene is identifiable, is not known by default in IoT and additional mechanisms for estimating sensors location might need to be employed.

The potential problem of low detection power of the FDR procedure is discussed in [20]. To address this critical question, a transformation of the p-value at each sensor is devised. However, a crucial assumption made is that the alternative hypothesis is known at the sensors [20], [21]. Otherwise, if the distribution under $\mathcal{H}_1$ is unknown, the transformation cannot be applied.

Furthermore, the distribution under the alternative hypothesis is implicitly assumed to remain constant (unchanged) over time.

In contrast to Ermis and Saligrama, we do not make any assumptions about the distribution of the observations under the alternative hypothesis nor about their distribution under the null hypothesis. In addition, the distribution under $\mathcal{H}_1$ (as well as under $\mathcal{H}_0$) can be different at each node. In effect, the local detectors used in our framework can on one hand learn the nominal conditions at each location and can, on the other hand, detect a change in the learnt conditions. The detection of a change in the nominal conditions does not require learning a priori the alternative distribution and in fact the alternative distribution can change over time. Such a change in the distribution of the observations under $\mathcal{H}_1$ does not have any relevant impact on the performance of the local detector in our framework.

Traditional distribution detection problems are tackled in [28], [29], [32], [33]. A *single-target* problem is examined in [28], [29] and [32], [33]. The primary difference between these two groups of studies is that in the former the target is within the sensing range of all nodes, whereas in the latter the target is within the sensing range of only a small fraction of the network nodes. In both cases, the traditional detection setup is assumed, where a single phenomenon but not multiple phenomena can be observed. Moreover, a specific signal, propagation and network models are studied. In particular, the nominal conditions are characterised by noise that obeys a standard normal distribution, $\mathcal{N}(0, 1)$, and the signal is modelled through a distance-dependent amplitude (an isotropic attenuation model is used). Furthermore, the sensors are uniformly distributed and the target is placed in the centre of the network. Although such assumptions are common, they determine the analytic derivations and parameter design. The system-level false alarm probability as well as detection probability are derived for this detection problem of Gaussian random variables. The distribution of the p-values under $\mathcal{H}_1$, for instance, is given assuming that when a target is present the signal follows $\mathcal{N}(\phi, 1)$, where ϕ is the received signal amplitude. For conditions different from the aforementioned the analytic derivations as well as parameter design and choice must be conducted anew. Hence, the devised solution is mainly relevant for the specific single-target scenario studied.

The radar signal processing problem of detecting multiple targets in clutter and noise problem is field-specific too. Moreover, a particular (Swerling I type of target) model in such radar scenarios is assumed. Thereby, the limited application of the provided solution in [30], [31].

Contrary to such application-specific and model-specific design, our framework is not limited to a particular application or model but allows for accommodating a large set of scenarios (single- and multi-object) and addressing a

variety of conditions (different empirical distributions under the null as well as alternative hypotheses).

In summary, earlier studies in wireless sensor networks that have adopted the FDR control procedure focus on the problem of decreasing the communication cost due to energy constraints typical for such networks. Consequently, Ermis and Saligrama develop a distributed FDR control procedure, which is later adopted by Ray and Varshney. In their work, the authors address the traditional set-up, where a target (or a set of targets) emits a signal. The hypothesis is of available knowledge either of the distributions of the observations under the null and alternative hypotheses or of specific models for the emitted signal(s) and signal attenuation. Contrary to such assumptions, in our work we do not make any assumptions about the observed target(s) but our statistical approach is applicable to any phenomena that can be sampled and modelled through an EDF. This makes the statistical framework versatile and applicable to large-scale IoT scenarios with a massive number of sensors. We also note that it is straightforward to incorporate the distributed FDR concept (**Algorithm** 6) into the statistical framework developed in our work.

Chapter 9

Conclusion

A statistical inference paradigm for Internet of Things scenarios, which are typically composed by a massive number of sensor nodes, is studied. We consider a traditional distributed scenario in which sensors make local observations about monitored phenomena and supply a central node (fusion centre) with test statistics. The fusion centre is responsible for making a global decision about the state of the network: departure from nominal conditions and/or localisation of events within the network. The statistical approach examined in the thesis combines non-parametric local detection of events and a multiple hypotheses testing procedure that controls error rates.

The main observation from the extensive simulation experiments is that the power of the statistical inference paradigm to detect occurrence of phenomena is primarily determined by the power of the local detectors. Other factors that can impact its power are the number of sensors and the proportion of true null hypotheses. On the other hand, the power of the non-parametric local detector employed in our work depends heavily on sample size and signal-to-noise ratio.

For a large number of sensors, demanding local conditions (probability models that resemble each other under the null and alternative hypotheses) and several sensors observing departure from nominal conditions, the power of the proposed statistical framework is large (above 90%). Nevertheless, when only few events occur (a small fraction of sensors observe a new phenomena), it is challenging for the fusion centre to detect them all (median power in the order of 75% for all but one of the examined cases). To understand and address this problem we analysed the simulation results, which brought insights into the performance of the proposed statistical inference approach as well as ways of improving it.

Future research prospects. Improving the performance of the local detector seems the most relevant next step as the overall power of the proposed statistical inference approach depends largely on the detection power of the sensors. Another idea for improving the detection probability at the fusion centre is characterisation of sensor performance: censoring or weighting of test statistics can be introduced based on the reliability of the nodes. Exploring correlation between sensor observations as well as clustering and collaboration between adjacent sensor nodes could potentially lead to improved power too (at the expense of increased communication complexity).

Design of decision rules at the fusion centre based on the obtained test statistics is the next research step within the statistical scope of our work.

Another relevant and interesting research task is to tackle the implications of the wireless environment – among them transmission loss and delay as well as bandwidth limitations can have an immediate impact on the attained power by the FDR control procedure.

Questions, which were intentionally left aside are the inherent energy constraints of sensor networks, the communication paradigm most suitable for large-scale networks as well as the secure transmission of information. The main reason not to tackle these relevant for IoT design choices is that the first step of this research is to verify the proposed concept and analyse the phenomena underlying the statistical performance without the impact of other factors. Nevertheless, in the broader scope of IoT the secure and energy-preserving communication of information needs to be addressed too.

Bibliography

[1] Anderson, T. W., Darling, D. A., "Asymptotic theory of certain "Goodness of Fit" criteria based on stochastic processes," *The Annals of Mathematical Statistics*, vol. 23, no. 2, pp. 193–212, 1952.

[2] Anderson, T. W., Darling, D. A., "A test of goodness of fit," *Journal of the American Statistical Association*, vol. 49, no. 268, pp. 765–769, 1954.

[3] Pettit, A. N., "A two-sample Anderson–Darling rank statistic," *Journal Biometrika*, Oxford University Press, Biometrika Trust, vol. 63, no. 1, pp. 161–168, 1976.

[4] Stephens, M., "The Anderson Darling statistic," *Technical Report No. 39*, Oct 31, 1979.

[5] Scholz, F. W., Stephens, M. A.,"K-Sample Anderson-Darling tests," *Journal of the American Statistical Association*, vol. 82, no. 399, pp. 918–924, 1987.

[6] Hall, P., and Wilson, S. R., "Two guidelines for bootstrap hypothesis testing," *Journal Biometrics*, Wiley, International Biometric Society, vol. 47, no. 2, pp. 757–762, 1991.

[7] Benjamini, Y. and Hochberg, Y., "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Jornal Royal Statistical Society*, vol. 57, no. 1, pp. 289–300, 1995.

[8] Lehmann, E. L. and Romano, J. P., *Testing statistical hypotheses*, Springer, Chapter 9, 2005.

[9] Goeman, J. J. and Solari, A.,"Multiple hypothesis testing in genomics," *Journal Statistics in Medicine*, Wiley, vol. 33, no. 11, pp. 1946–1978, 2014.

[10] Dudoit, S., Shaffer, J., and Boldrick, J., "Multiple hypothesis testing in microarray experiments," *Journal Statistical Science*, Wiley, vol. 18, no. 1, pp. 71–103, 2003. http://www.jstor.org/stable/3182872.

[11] Efron, B., and Tibshirani, R. J. *An introduction to the bootstrap.* Chapman & Hall, CRC Monographs on Statistics & Applied Probability, 1994.

[12] Efron, B., *Large-scale inference. Empirical Bayes methods for estimation, testing, and prediction.* Cambridge University Press, 2013.

[13] Efron, B., "Comment: Microarrays, empirical Bayes and the two-groups model," *Journal Statistical Science*, Institute of Math Stat., vol. 23, no. 1, pp. 1–22, 2008.

[14] Benjamini, Y., "Comment: Microarrays, empirical Bayes and the two-groups model," *Journal Statistical Science*, Institute of Math Stat., vol. 23, no. 1, pp. 23–28, 2008.

[15] Cai, T. T., "Comment: Microarrays, empirical Bayes and the two-groups model," *Journal Statistical Science*, Institute of Math Stat., vol. 23, no. 1, pp. 29–33, 2008.

[16] Morris, C. N., "Comment: Microarrays, empirical Bayes and the two-groups model," *Journal Statistical Science*, Institute of Math Stat., vol. 23, no. 1, pp. 34–40, 2008.

[17] Benjamini, Y. and Hochberg, Y., "On the adaptive control of the false discovery rate in multiple testing with independent statistics," *Journal Educational and Behavioral Statistics*, vol. 25, no. 1, pp. 60–83, 2000.

[18] Benjamini, Y. and Yekuteli, D., "The control of the false discivery rate in multiple testing under dependency," *The Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, 2001.

[19] Kay, S. M., *Fundamentals of statistical signal processing.* Prentice Hall, 1993.

[20] Ermis, E. B. and Saligrama, V., "Adaptive statistical sampling methods for decentralized estimation and detection of localized phenomena," *IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'05),* 2005, vol. 5, no. 4, pp. V-1045–V-1048.

[21] Ermis, E. B. and Saligrama, V., "Adaptive statistical sampling methods for decentralized estimation and detection of localized phenomena," *IEEE 4rth Int. Symp. Information Processing in Sensor Networks (IPSN'05),* 2005, pp. 1–8.

[22] Ermis, E. B. and Saligrama, V., "Search and discovery in an uncertain networked world," *IEEE Signal Processing Mag.*, vol. 23, no. 4, pp. 107–118, 2006.

[23] Ermis, E. B. and Saligrama, V., "Detection and localization in sensor networks using distributed FDR," *IEEE 40th Annual Conf. Information Sciences and Systems*, 2006, pp. 699–704.

[24] Ermis, E. B. and Saligrama, V., "Robust distributed detection with limited range sensors," *IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'07)*, 2007, vol. 2, pp. II-1009–II-1012.

[25] Ermis, E. B. and Saligrama, V., "Dynamic thresholding for distributed multiple hypotheses testing," *IEEE 14th Workshop Statistical Signal Processing*, 2007, pp. 675–679.

[26] Ermis, E. B. and Saligrama, V., "Distributed detection in sensor networks with limited range multimodal sensors," *IEEE Trans. Signal Processing*, vol. 2, pp. 843–858, 2010.

[27] Rago, C., Willett, P. and Bar-Shalom, Y., "Censoring sensors: a low-communication-rate scheme for distributed detection," *IEEE Trans. Aerospace and electronic systems*, vol. 32, no. 2, pp. 554–568, 1996.

[28] Ray, P., Varshney, P. K. and and Niu, R., "A novel framework for the network-wide distributed detection problem," *IEEE 10th Int. Conf. Information Fusion*, 2007, pp. 1–8.

[29] Ray, P. and Varshney, P. K., "Distributed detection in wireless sensor networks using dynamic sensor thresholds," *Int. J. Distributed Sensor Networks*, vol. 4, no. 1, pp. 4–11, Jan. 2008.

[30] Ray, P. and Varshney, P. K., "A false discovery rate based detector for detection of targets in clutter and noise," *IEEE 10th Int. Conf. Information Fusion*, 2008, pp. 1–6.

[31] Ray, P. and Varshney, P. K., "Radar target detection framework based on false discovery rate," *IEEE Trans. Aerospace and electronic systems*, vol. 47, no. 2, pp. 1277–1291, 2011.

[32] Ray, P. and Varshney, P. K., "False discovery rate based sensor rules for the network-wide distributed detection problem," *IEEE Trans. Aerospace and electronic systems*, vol. 47, no. 3, pp. 1785–1799, 2011.

[33] Vempaty, A., Ray, P. and Varshney, P. K., "False discovery rate based distributed detection in the presence of Byzantines," *IEEE Trans. Aerospace and electronic systems*, vol. 50, no. 3, pp. 1826–1840, 2014.

[34] Chair, Z. and Varshney, P. K., "Optimal data fusion in multiple sensor detection systems," *IEEE Trans. Aerospace and electronic systems*, vol. 22, no. 1, pp. 98–101, 1986.

[35] Rohling, H., "Radar CFAR thresholding in clutter and multiple target situations," *IEEE Trans. Aerospace and electronic systems*, vol. AES-19, no. 4, pp. 608–621, July 1983.

[36] Gelman, A., Carlin J. B., Stern, H. S., Danson, D. B., Vehtari, A. and Rubin, D. B. *Bayesian data analysis*, CRC Press, 3rd edition, 2014.

[37] Varsheny, P. K., *Distributed Detection and Data Fusion*, Springer, 1997.

[38] Efron, B., "Size, power and false discovery rate," *The annals of Statistics*, vol. 35, no. 4, pp. 1351–1377, 2007.

[39] Pawitan, Y., Michiels S., Koscielny S., Gusnanto A. and Ploner A. "False discovery rate, sensitivity and sample size for microarray studies," *Bioinformatics*, vol. 21, no. 13, pp. 3017–3025, 2005.

[40] M. Gölz, V. Koivunen, and A. Zoubir, "Nonparametric detection using empirical distributions and the bootstrap," *25th European Signal Processing Conference* (EUSIPCO), 2017, pp. 1450-1454.

[41] Akyildiz, I. F., SU, W., Sankarasubramaniam, Y., Cayirci, E., "Wireless sensor networks: a survey," *Elsevier Computer Networks*, vol. 38, no. 4, pp. 393–422, 1952.

[42] Yick, J., Mukherjee, B., Ghosal, D."Wireless sensor network survey," *Elsevier Computer Networks*, vol. 52, no. 12, pp. 2292–2330, 2008.

[43] Chamberland, J. F., and Veeravalli, V. V., "Wireless sensors in distributed detection applications," *IEEE Signal Processing Magazine*, vol. 24, no. 3, pp. 16–25, May 2007.

[44] Afzala, B., Alvi, S. A., Shahb, G. A., Mahmoodb W., "Energy efficient context aware traffic scheduling for IoT applications," *Elsevier Ad Hoc Networks*, vol. 62, no. 12, pp. 101–115, 2017.

[45] Husnain, H., Sherazi, R., Grieco, L. G., Boggia, G., "A comprehensive review on energy harvesting MAC protocols in WSNs: challenges and tradeoffs," *Elsevier Ad Hoc Networks*, vol. 71, no. 12, pp. 117–134, 2018.

[46] Shridhar, V. S., "Tata communications' countrywide Internet of Things will manage the chaos in India's booming cities," *IEEE Spectrum*, January, 2019.

[47] Chaen, B., Tong, L., and Varsheny, P. K., "Channel-aware distributed detection in wireless sensor networks,", *IEEE Signal Processing Magazine*, vol. 23, no. 4, pp. 16–26, July 2006.

Appendix A

FDR sensitivity results

A.1 Total number of hypotheses m

Table A.1: Effect of m on FDR power when detecting a change in the mean of a Gaussian. $\mathrm{SNR}_{dB} = -6$ dB, FDR threshold $q^* = 0.05$ and proportion of true null hypotheses $\pi_0 = 0.1$. 95 % BCIs for the power and theoretical $\mathbb{E}_{\mathrm{th}}[Q]$ and simulated $\mathbb{E}_{\mathrm{sim}}[Q]$ FDR are reported.

m	$\mathbb{E}_{\mathrm{sim}}[Q]$	$\mathbb{E}_{\mathrm{th}}[Q]$	power	n
10^1	$[.0041, .0046]$	≤ 0.005	$[.1677, .1699]$	10
10^2	$[.0047, .0050]$		$[.1169, .1173]$	
10^3	$[.0050]$		$[.1104, .1107]$	
10^1	.0050	≤ 0.05	.9986	100
10^3	$[.00499, .05000]$	≤ 0.05	.9986	
10^4	$[.00499, .05000]$		.9986	
$\{10, 10^4\}$	0.0050	≤ 0.05	1	1000

Table A.2: Effect of m on FDR power when detecting a difference in the variance of two zero-mean Gaussians: $\mathrm{SNR}_{dB} \approx -3.5$ dB, $q^* = 0.05$ and $\pi_0 = 0.5$.

m	$\mathbb{E}_{\mathrm{sim}}[Q]$	$\mathbb{E}_{\mathrm{th}}[Q]$	power	n
10^1	0.0229	≤ 0.025	$[0.0422, 0.0434]$	10
10^2	0.0224		$[0.0069, 0.0075]$	
10^3	0.0222		$[0.0010, 0.0011]$	
10^3	0.0248		$[0.9605, 0.9617]$	100

A.2 Proportion of true null hypotheses π_0

Table A.3: Impact of the proportion of true null hypotheses π_0 on FDR power. A change in the mean is detected under $\mathrm{SNR}_{dB} = -6$ dB; $q^* = 0.05$ and $\pi_0 = 0.9$. 95 % bootstrapped CIs for the mean values are reported.

m	$\mathbb{E}_{\mathrm{sim}}[Q]$	$\mathbb{E}_{\mathrm{th}}[Q]$	power	n
10^1	$[.0415, .0439]$	≤ 0.045	$[.1108, .1149]$	10
10^2	$[.0403, .0427]$		$[.0339, .0347]$	
10^3	$[.0390, .0409]$		$[.0110, .0112]$	
10^1	$[.0443, .0461]$	≤ 0.045	$[.9854, .9868]$	100
10^2	$[.0448, .0451]$		$[.9862, .9863]$	
10^3	0.045	≤ 0.045	$[.9861, .9862]$	
10^1	0.045		1	1000
10^2	$[0.0448, 0.0451]$		1	
10^3	0.045		0.9862	

When the sampled data points are few, FDR power decreases whenever π_0 increases (compare Table A.3 to Table A.1); otherwise under sufficiently large sample size n, FDR power remains unaffected by the the proportion of true null hypothese π_0.

A.3 FDR threshold q^*

Table A.4: FDR performance results—95% BCI of the power and maximum and mean of the number of false discoveries—with two different thresholds q^* and for very small sample size $n = 10$ when $\text{SNR}_{dB} = -6$ dB. The proportion of true null hypotheses is $\pi_0 = 0.1$. The local hypothesis test is the z-test.

q^*	m	power	$\max(V)$	$\mathbb{E}[V]$
0.05	10^2	[.1169, .1173]	4	0.0636
	10^3	[.1104, .1107]	6	0.5119
0.2	10^2	[.4314, .4318]	7	0.8254
	10^3	[.4325, .4337]	23	7.9646

Table A.5: FDR performance results with two different q^* thresholds and for extremely low $\text{SNR}_{dB} \approx -30$dB. The total number of hypotheses is $m = 10^4$ and sample size $n = 10^3$. The local hypothesis test is the z-test. 95 % bootstrapped confidence intervals of the mean of the power are reported.

q^*	π_0	$\mathbb{E}_{\text{sim}}[Q]$	$\mathbb{E}_{\text{th}}[Q]$	power	$\max(V)$
0.05	0.1	.0042	$\leq .005$	[.0027, .0028]	4
	0.9	.0410	$\leq .045$	[.00026, .00029]	3
0.2	0.1	.0199	$\leq .020$	[.0896, .0900]	38
	0.9	.1443	$\leq .180$	[.0017, .0018]	17

In the scenarios studied in Table A.4 and Table A.5 larger (less stringent) FDR bound q^* improves FDR power but at the cost of an increase in the number of false discoveries V.

A.4 SNR conditions

Table A.6: Impact of SNR conditions on FDR performance with the z-test when FDR threshold is $q^* = 0.05$, $\pi_0 = 0.1$ and sample size is $n = 10^3$. When the SNR is extremely low, FDR power is close to zero. F denotes the distribution under nominal conditions and G denotes the conditions under the alternative hypothesis.

m	F	G	π_0	$\mathbb{E}_{\text{sim}}[Q]$	$\mathbb{E}_{\text{th}}[Q]$	p_d	SNR_{db}
$1 * 10^4$	$\mathcal{N}(0,1)$	$\mathcal{N}(0.5,1)$	0.5	.0250	$\leq .0250$	1	-6
			0.7	.0350	$\leq .0350$	1	
			0.9	.0450	$\leq .0450$	1	
			1.0	.0514	$\leq .0500$	–	
$1 * 10^3$	$\mathcal{N}(0,15^2)$	$\mathcal{N}(0.5,15^2)$	0.5	.0229	$\leq .0250$	0.00260	-30
$5 * 10^3$	$\mathcal{N}(0,15^2)$	$\mathcal{N}(0.5,15^2)$	0.7	.0273	$\leq .0350$	$4.94 * 10^{-4}$	
			0.9	.0400	$\leq .0450$	$2.87 * 10^{-4}$	
$1 * 10^4$	$\mathcal{N}(0,15^2)$	$\mathcal{N}(0.5,15^2)$	0.9	.0401	$\leq .0250$	0.00028	
$1 * 10^1$	$\mathcal{N}(0,15^2)$	$\mathcal{N}(0.5,15^2)$	0.1	.0052	$\leq .0050$	0.0497	
			0.9	.0443	$\leq .0455$	0.0400	

Appendix B

Power of the local detector

Table B.1: 95 % bootstrapped confidence intervals for the mean and median of the acceptance ratio of the non-parametric local detector when evaluating the null hypohtesis that ambient sample X and observation sample Y obey the same distribution F. The detector is tuned with the values reported in Table 7.4. The significance level is $\alpha = 0.05$.

F	G	mean	median
Exp(1)	$\mathcal{N}(1,1)$	[0.001400, 0.001500]	[0.001300, 0.001400]
$\mathcal{N}(1,1)$	Exp(1)	[0.034200, 0.035700]	[0.032300, 0.034100]
Exp(1)	Exp(1)	[0.948600, 0.949500]	[0.948900, 0.949800]
$\mathcal{N}(1,1)$	$\mathcal{N}(1,1)$	[0.948400, 0.949300]	[0.948500, 0.949700]
Ray(1)	$\mathcal{N}(1,1)$	[0.000854, 0.000903]	[0.000800, 0.000900]
$\mathcal{N}(1,1)$	Ray(1)	[0.000252, 0.000283]	[0.000200, 0.000200]
Ray(1)	Ray(1)	[0.948500, 0.949400]	[0.948600, 0.949600]

Appendix C

Local conditions

C.1 Empirical cumulative distribution functions for Central vs Non-central χ^2

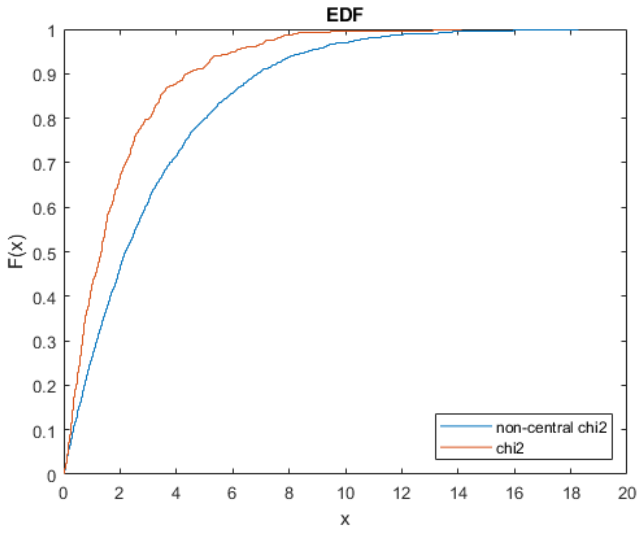

Figure C.1: EDF sampled from Noncentral $\chi^2(2)$ vs Central $\chi^2(2)$ ($n = 500$ and $l = 100$, respectively).

C.2 EDFs for Rician and Rayleigh

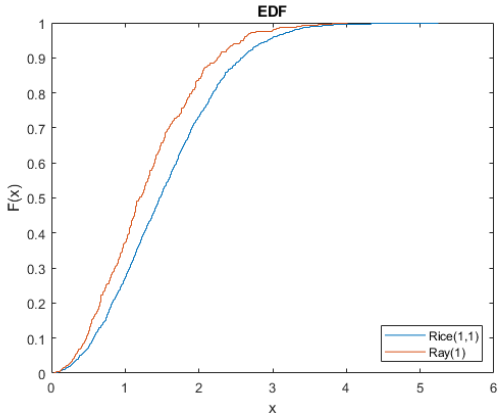

Figure C.2: EDF sampled from Rician ($n = 500$) and Rayleigh ($l = 100$) distributions.

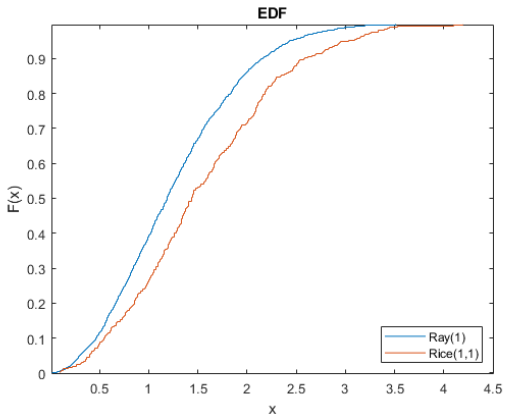

Figure C.3: EDF sampled from Rayleigh ($n = 500$) and Rician ($l = 100$) distributions.

C.3 EDFs for Rayleigh and Normal

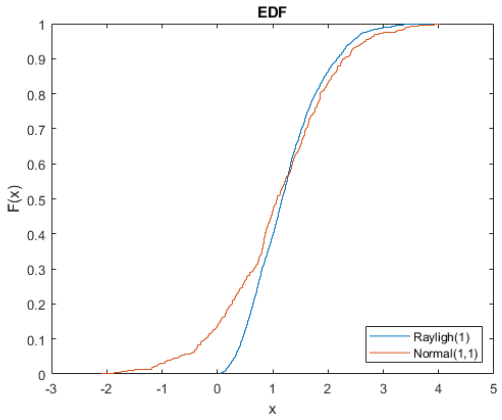

Figure C.4: EDF sampled from Rayleigh ($n = 500$) and Normal ($l = 100$) distributions.

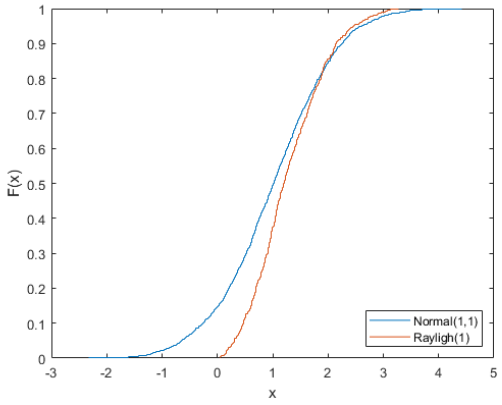

Figure C.5: EDF sampled from Normal ($n = 500$) and Rayleigh ($l = 100$) distributions.

C.4 EDFs for Normal and Exponential

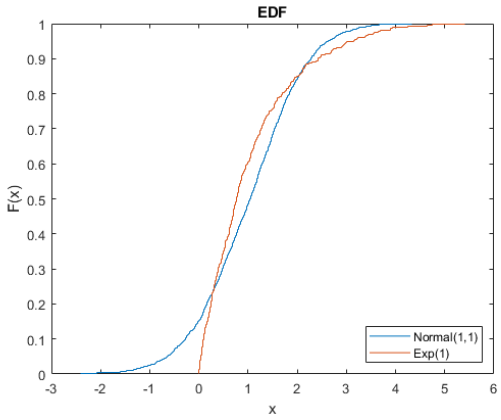

Figure C.6: EDF sampled from Normal ($n = 500$) and exponential ($l = 100$) distributions.

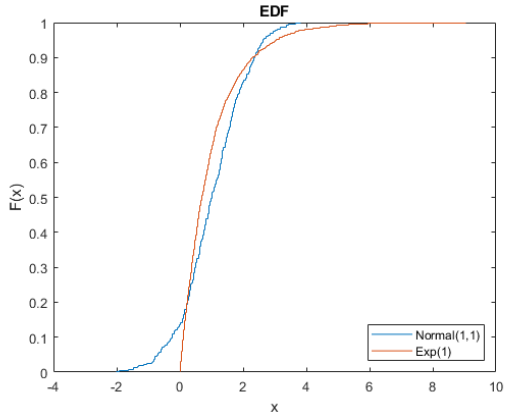

Figure C.7: EDF sampled from exponential ($n = 500$) and Normal ($l = 100$) distributions.